Wireless Communications Fundamentals

EB-553A TEXTBOOK 595-5929-02

WIRELESS COMMUNICATIONS FUNDAMENTALS, Textbook

Copyright © 2001 by Heathkit Company, Inc., Benton Harbor, Michigan, 49022. All Rights Reserved. Printed in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, electronic or mechanical, including photocopying, recording, storage in a data base or retrieval system, or otherwise, without the prior written permission of the publisher.

ISBN 0-87119-316-7

CONTENTS

Introduction iv

Course Objectives v

Course Outline vi

Section 1 — Communication Fundamentals
Chapter 1 — What is Communication? 1-1
Chapter 2 — Review of Complex Waves 2-1
Chapter 3 — Amplitude Modulation 3-1
Chapter 4 — Angle Modulation 4-1
Chapter 5 — The Communication System 5-1

Section 2 — Amplitude Modulation
Chapter 6 — AM Basics 6-1
Chapter 7 — Suppressed Carrier AM 7-1
Chapter 8 — Single Sideband 8-1

Section 3 — AM Receivers
Chapter 9 — AM Detectors 9-1
Chapter 10 — Tuner RF and Superheterodyne Receivers 10-1
Chapter 11 — Receiver Analysis 11-1

Section 4 — Angle Modulation
Chapter 12 — Angle Modulation 12-1
Chapter 13 — Demodulators 13-1
Chapter 14 — Receivers 14-1
Chapter 15 — FM Transceivers 15-1

Section 5 — Pulse Modulation
Chapter 16 — Analog Pulse Modulation 16-1
Chapter 17 — Digital Pulse Modulation 17-1
Chapter 18 — Multiplexing 18-1

Section 6 — Antennas
Chapter 19 — Transmission Lines 19-1
Chapter 20 — Antennas 20-1

Glossary G-1

Index I-1

INTRODUCTION

This course introduces the fundamental building blocks of electronic communications systems. It examines information, modulation, complex waves, transmitters, receivers, and communication channels.

You will analyze methods of generating amplitude-modulated signals, and examine the characteristics of those AM signals in detail. A newer type of AM, suppressed carrier AM, is examined in several ways. One form of it, single sideband, is found to offer a reasonable compromise between complexity and performance. How AM is used to transmit digital signals is also discussed.

You will study circuits and techniques that are used to receiver and detect amplitude-modulated signals. A chapter also investigates angle modulation, and demodulation, to augment your knowledge of modulation, transmission, and receivers. To combine these concepts, we will examine schematic diagrams of two typical angle modulation devices. Sampling is covered in depth to insure you understand how it is applied to pulse modulation.

In the last two chapters, you will learn about transmission lines and antennas. There are many types of transmission lines available, depending upon the transmitter power level, frequency, and length of line required. Similarly, antennas come in many shapes and sizes based upon intended frequency and purpose.

COURSE OBJECTIVES

When you complete this course you will be able to:

- Explain the characteristics, advantages, and disadvantages of amplitude, angle, and pulse modulation.
- Break down complex waveforms to their component parts.
- Explain the operation of AM and FM transmitters and receivers.
- Analyze and explain the construction of basic communication circuits.
- Explain the differences between frequency and phase modulation.
- Explain how frequency- and time-division multiplexing are used to increase channel capacity.
- State the type, characteristics, and uses of transmission lines.
- List and explain the basic types of antennas.

COURSE OUTLINE

Introduction

Section 1 — Communication Fundamentals

Chapter 1 – What is Communication?
- Introduction
- Objectives
- What is Information?
- The Capacity of a Communication Channel
  - What is Bandwidth?
  - Shannon's Law
- Information Transfer
  - Modulation
  - Types of Modulation
- Summary
- Self-Test Review

Chapter 2 – Review of Complex Waves
- Introduction
- Objectives
- Sine Wave
- Square Wave
- Sawtooth Wave
- Fourier Analysis
- Bandwidth of Complex Signals
- Summary
- Self-Test Review

Chapter 3 – Amplitude Modulation
- Introduction
- Objectives
- The AM Waveform
- Percent of Modulation
- Sidebands
- Bandwidth
- Summary
- Self-Test Review

Chapter 4 – Angle Modulation
- Introduction
- Objectives
- Frequency Modulation
- Phase Modulation
- Modulation Index

Sidebands
Bandwidth
Angle Modulation and Pulses
Summary
Self-Test Review

Chapter 5 – The Communication System
Introduction
Objectives
Input/Output Devices
Transmitter
Channel
Noise
Distortion
Loss
Receiver
Summary
Self-Test Review

Section 2 — Amplitude Modulation

Chapter 6 – AM Basics
Introduction
Objectives
Creating Amplitude Modulation
The Diode Modulator
The Absorption Modulator
The Collector Modulator
Series Modulator
Progressive Series Modulator
AM Characteristics
Power and Efficiency
Bandwidth
Phase Relationships
Noise and Distortion
Amplitude Modulation, Pulses, and Noise
Summary
Self-Test Review

Chapter 7 – Suppressed Carrier AM
Introduction
Objectives
Balanced Modulators
Twin-FET Balanced Modulator
Diode-Ring Balanced Modulator
IC Balanced Modulators

Disadvantages of Suppressed Carrier AM
Summary
Self-Test Review

Chapter 8 – Single Sideband
Introduction
Objectives
Filter Method SSB Transmitter
Sideband Filters
Frequency Conversion
Sideband Selection
Phasing Method of SSB
Linear Amplifiers
SSB Power Ratings
Summary
Self-Test Review

Section 3 — AM Receivers

Chapter 9 – AM Detectors
Introduction
Objectives
The Diode Detector
The Transistor Detector
The Synchronous Detector
The Product Detector
Summary
Self-Test Review

Chapter 10 – Tuned RF and Superheterodyne Receivers
Introduction
Objectives
Tuned RF Receiver
Sensitivity
Selectivity
Superheterodyne Receivers
Block Diagram
Frequency Conversion Circuits
Intermediate Frequency Selection
IF Amplifiers
Automatic Gain Control
Double Conversion Receivers
Other Applications
Summary
Self-Test Review

Chapter 11 – Receiver Analysis
- Introduction
- Objectives
- Receiver Specifications
  - Sensitivity
  - Selectivity
  - Image Rejection
  - IF Rejection
  - Dynamic Range
  - Noise Figure
  - Overall Gain
  - Distortion
  - AGC Performance
- Analysis of a Communications Receiver
  - Specifications
  - Circuit Analysis
  - AGC Circuit
- Summary
- Self-Test Review

Section 4 — Angle Modulation

Chapter 12 – Angle Modulation
- Introduction
- Objectives
- Reactance Modulator
- Varactor Modulator
- Phase Modulator
- A Complete Transmitter
- Summary
- Self-Test Review

Chapter 13 – Demodulators
- Introduction
- Objectives
- Slope Detector
- Double-Tuned Detector
- Foster-Seeley Discriminator
- Ratio Detector
- Phase-Locked-Loop Demodulator
- Summary
- Self-Test Review

Chapter 14 – Receivers
Introduction
Objectives
FM Broadcast Receiver
AM/FM Broadcast Tuner
Summary
Self-Test Review

Chapter 15 – FM Transceivers
Introduction
Objectives
Schematic Analysis
Summary
Self-Test Review

Section 5 — Pulse Modulation

Chapter 16 – Analog Pulse Modulation
Introduction
Objectives
Sampling
Pulse Amplitude Modulation
Pulse Time Modulation
Pulse Duration Modulation
Pulse Position Modulation
Summary
Self-Test Review

Chapter 17 – Digital Pulse Modulation
Introduction
Objectives
Quantization
Quantization Noise
Variable-Level Quantization
More About Sampling
Coding
PCM Transmitters and Receivers
Other DPM Systems
Analog and Digital Pulse Modulation Compared
Summary
Self-Test Review

Chapter 18 – Multiplexing
Introduction
Objectives
The 4 kHz Channel: A Basic Building Block
Space-Division Multiplexing
Advantages and Disadvantages
Frequency-Division Multiplexing
Advantages and Disadvantages
Time-Division Multiplexing
Practical TDM-PCM Systems
Advantages and Disadvantages
Summary
Self-Test Review

Section 6 — Antennas

Chapter 19 – Transmission Lines
Introduction
Objectives
The Balanced Transmission Line
Standing Waves
Standing Wave Ratio
Transmission Line Resonant Circuits
Coaxial Transmission Lines
Comparison of Coaxial and Balanced Lines
Waveguides
SWR Measurement
Summary
Self-Test Review

Chapter 20 – Antennas
Introduction
Objectives
Antenna Principles
Antenna Radiation
The Dipole Antenna
The Vertical Antenna
Horn Arrays
Antenna Arrays
Parasitic Arrays
Driven Arrays
Summary
Self-Test Review

Section 1

Communication Fundamentals

CHAPTER 1 — What is Communication?

INTRODUCTION

This chapter begins to discuss the fundamental building blocks of electronic communications systems.

Communication is a process by which information is exchanged. Letters, newspapers, and conversation are all forms of communications. Electronic communications systems use electronic technology to transmit and receive information. Radio, television, and the telephone system are all examples of electronic communications systems. The purpose of any communications system is to convey or transfer information from one point to another.

OBJECTIVES

When you complete this chapter, you will be able to:

- Define communication, information, modulation, and carrier.
- List at least two reasons why information cannot be transmitted at its original frequency.
- Name the three basic types of modulation.
- Describe how information is measured.

WHAT IS INFORMATION?

Symbols are used to express information. The letters and words on this page are symbols which represent information. Information can be measured. While information theory is very complex, in general we can describe information in terms of the relationship between two elements: the amount of information contained in a symbol and the predictability of the symbol. Information is knowledge you did not previously have. If you already know the message, it has no information content. That's why unpredictability and high information content go hand in hand. Let's take an example.

If I send to you a six-letter English language word consisting of the letters F R I D A, you expect the next letter to be Y. It might not be, but it probably is, because FRIDAY is a common English word which contains six letters begins with the five letters you already know. In most dictionaries, once you know the first four letters of this particular message, only two or three words exist that are possibilities for the correct solution. We could say then that the first five letters of FRIDAY contained more information than the last one, because we could guess the last one quite accurately once we knew the first five.

In fact, each succeeding letter in the message conveys less information than the ones before it, because each additional letter reduces the possible number of correct words that the message could be. The first letter reduces the choices to all six-letter English words starting with "F," the second letter to all those words starting with "FR," and so forth. The relationship between information content and predictability for this message is shown in Figure 1-1.

Figure 1-1 shows that as each letter's information content goes down, it becomes more predictable. This feature of language is the basis of crossword puzzles. It is called redundancy.

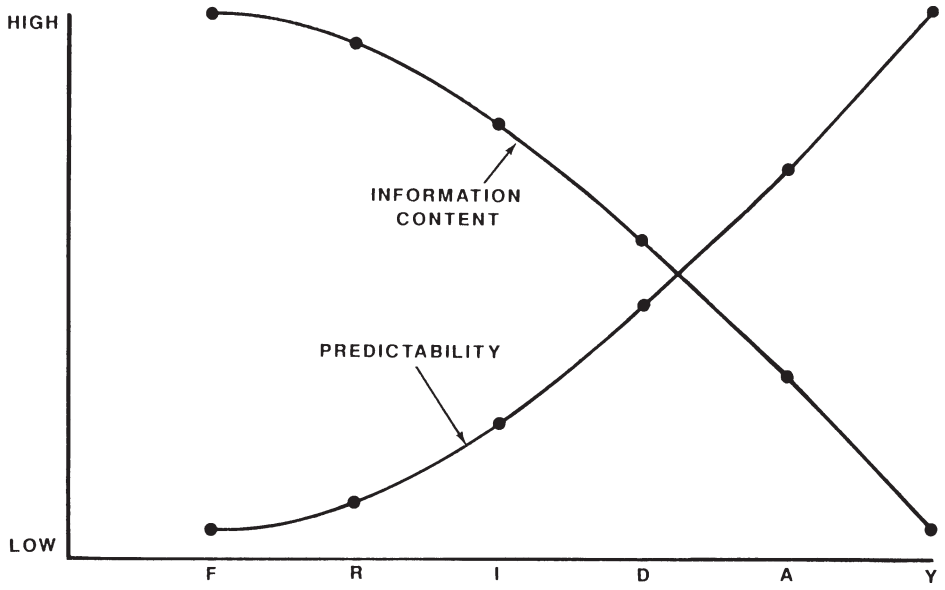


Figure 1-1

Graph comparing information content and predictability for the example given in the text.

Human language is very redundant. It has been estimated that as many as 75% of our spoken words are redundant, and so speech is said to have an information content of about 25%. You know this intuitively; if you miss a word or two during a speech or television broadcast, you can usually guess correctly what was said. Not too much information was lost due to a few missing words. This is especially true if the missed words do not occur all at once.

Computers, on the other hand, use digital data to converse. Data is not very redundant, so it is said to have a high information content. It is practically impossible to predict what the next bit in a data stream will be, even knowing the one that was just received. And it is usually much more serious to lose a single number than it is a single word.

Suppose that a spreadsheet containing your company's pay records is being transmitted between computers. The number opposite your name that represents your annual salary should be 27500, but the third number is lost during transmission. The receiving computer gets a spreadsheet that says your annual pay is 2700. Rather a significant difference! Because of its high information content and low redundancy, special care must be taken in transmitting digital data to guard against lost or altered data, and to detect if problems do occur. Handling of digital data is discussed in detail later in the course.

What about the information in the familiar sine wave, which is widely used in communications? If we know its amplitude, frequency, and period, the sine wave can be accurately predicted at any future point. If we can know everything about the future sine wave based upon the present, then it does not convey any information. We say that its information content is zero.

If information can have a value of zero, it must also be able to have values greater than zero, or we could never communicate. This is exactly so. Information can be measured, just as volts, amperes, and ohms can be. Measuring information is very important in electronic communications systems.

THE CAPACITY OF A COMMUNICATIONS CHANNEL

A communications channel is like a water pipe. The water flowing in the pipe represents information flowing in a communications channel. The capacity of the pipe to carry water depends upon the pressure of the water and the diameter of the pipe. Exactly the same is true of a communications channel. The faster we push information into the channel, and the wider the channel is, the more information it will transmit.

You cannot increase the pipe diameter whenever you want; once you have bought the pipe, the diameter is fixed. The only way to increase the flow of water through the pipe is to increase the pressure. If you keep increasing the pressure without limit, the pipe breaks! So you can see that the pipe has a maximum capacity to transmit water, and that capacity depends upon the size of the pipe and the water pressure.

Communication channels, like pipes, come in certain sizes. The size of a communications channel is measured in bandwidth, which has the dimensions of Hertz. This is equivalent to the pipe diameter. The speed with which we send digital information over the channel is measured in bits per second, it is like the water pressure. Just as with the pipe, there is a maximum that the channel can carry.

What is Bandwidth?

The bandwidth of a communication path is measured in Hertz, and is calculated by subtracting the minimum frequency that will pass through the channel from the maximum frequency that will pass. Said differently, bandwidth tells you the width of the band of frequencies

that a circuit will pass from input to output. It does not tell you anything about the placement of those frequencies in the spectrum. The wider the bandwidth, the more information the channel can carry. The narrower the bandwidth, the less information it can carry.

Consider the circuit response curve shown in Figure 1-2. What is its bandwidth? It all depends upon where you decide to measure, doesn't it? To avoid ambiguous measurements, it is important to decide in advance on what signal level to use for measuring bandwidth. Unless otherwise specified, bandwidth normally is measured at points 3 dB below the maximum level transmitted through the channel. Figure 1-3 illustrates this. The average maximum signal level is designated the 0 dB level. We then find the frequencies at which the received signal has fallen to 3 dB below that level; then, we subtract the lower frequency from the higher. The result is the channel bandwidth.

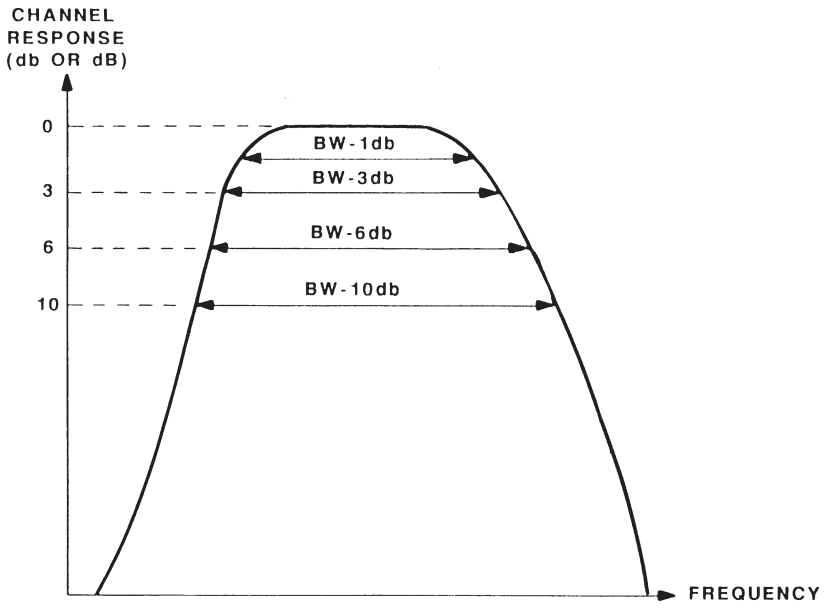


Figure 1-2

Channel response curve and some bandwidth measurements.

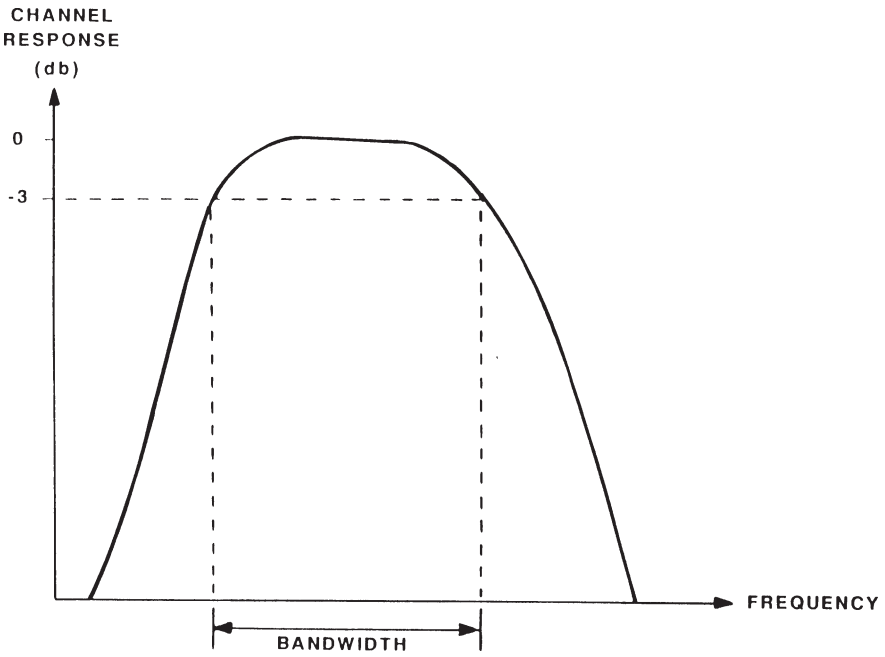


Figure 1-3

Channel response curve showing bandwidth measured at the 3-dB points (standard measurement).

Shannon's Law

In 1948, Claude Shannon proved that the maximum capacity of a communication channel using an unpredictable (that is, non-redundant) signal can be calculated using the equation:

$$C = W \log_2\left(1 + \frac{S}{N}\right)$$

where:

C = channel capacity in bits per second

W = channel bandwidth in Hertz (Hz)

$\frac{S}{N}$ = signal-to-noise ratio of the channel

All of the quantities on the right side of the equation are measurable, and are usually known. Therefore, we can calculate the maximum amount of information that can be sent over that channel.

Let's see what the maximum information capacity of a "typical" telephone circuit might be. Suppose our typical circuit has a bandwidth of 3100 Hz, and that its signal-to-noise ratio is 20 decibels (which is a ratio of 100:1). The logarithm in Shannon's equation is base 2; you may be more accustomed to a base 10 system. We can convert by multiplying by 3.32, as follows:

$$C = W \times 3.32 \times \log\left(1 + \frac{S}{N}\right)$$

$$C = 3100 \times 3.32 \times \log\left(1 + \frac{100}{1}\right)$$

C = 20,628 bits per second.

As you see, our "typical" telephone circuit cannot transmit data at a rate faster than 20,628 bits per second. Unlike the pipe, the circuit will not burst if we try to send data faster than this rate. The circuit simply won't accept any more signals, and information will be lost. The only way to improve its performance would be to change a fundamental characteristic of the circuit: its bandwidth or its signal-to-noise ratio.

For these two main reasons, practical communication systems usually operate at speeds much less than the Shannon limit, especially when they operate over telephone lines. First, significantly changing either the bandwidth or the signal-to-noise ratio of telephone circuits is difficult. Second, coding of the data becomes more complex as you approach the Shannon limit. Either of these approaches increases costs dramatically. Later in the course, we'll look at trade-offs between cost and speed, and how they affect communication channels.

For now, the important thing to remember is that information is a measurable quantity, and how much of it can be sent over a communication channel depends on the characteristics of the channel.

INFORMATION TRANSFER

Our most fundamental methods of transferring information are speech and writing. Communication of the written word developed from hand-carried letters and newspapers to the mail system, telegraph, and now electronic mail. Spoken communication evolved from face-to-face contact into telephone and radio communication. All of these steps were taken in an effort to increase the communication distance and speed.

The problem with transmitting speech is that the distance you can communicate is limited to how far your voice can be heard. If two people try to communicate over the same path at the same time, their voices interfere with one another. With written communication, you have to physically transport the written page from writer to reader. Until recently, this could be done only at the speed that a man could travel. The advent of the telegraph, telephone, and radio overcame these disadvantages.

Telegraph converted written messages into electrical signals and sent them long distances over wires at the speed of light. At the receiver, the signals were converted to sound and written down again. Soon, machines were developed that automated the conversion back to writing, speeding the information transfer even more. The telephone followed about thirty years later, giving people the ability to conduct conversations that spanned first hundreds, and then thousands of miles. Instant communication over long distances had become a reality.

However, a major drawback still existed. Each conversation or written message required its own pair of wires, otherwise it would interfere with other conversations or messages. Radio overcame this obstacle. Using radio, information is shifted upwards in frequency from the voice band to radio frequencies. A different conversation can be held over each of the several radio frequencies; the listener chooses the correct one by tuning to the proper frequency. Radio provides a separate path for each conversation, so they no longer interfere with one another.

Radio also makes information transfer more efficient. Radio waves propagate for long distances, without wires. Voice frequencies and written messages don't. Using radio, much of the clutter of wires needed by telegraph and telephone systems could be eliminated, along with the cost of installing and maintaining them. Systems could even be built using radio technology, interconnected by cables, thereby avoiding the problems of on-air interference.

Radio solved the problem of having everyone trying to communicate on voice frequencies at the same time. The radio signal used to send information is called a carrier. The process of adding information to a carrier is called modulation; the process of recovering the information from the carrier is called demodulation. Modulation is the key to electronic communication, and will form a large part of this course.

> The radio signal used to send information is called a *carrier*. The process of adding information to a carrier is called *modulation*.

Modulation

In the process of modulation, some characteristic of a high frequency sine wave is varied in accordance with the information or modulating signal. This signal may be an audio waveform, a digital pulse train, a television picture, or any other form of information. The important consideration is that it is shifted to a higher frequency for efficient transmission.

The high frequency sine wave is called the carrier. The mathematical expression for an unmodulated sine wave or carrier is:

$$e = A \sin (\omega t + \phi)$$

Where

e = instantaneous value of the wave (voltage or current)
A = maximum amplitude
ω = angular velocity ($2 \pi f$)
ϕ = phase angle

This equation shows that there are three characteristics of the wave that can be varied or modulated. These are: amplitude (A), angular velocity or frequency (ω), and phase angle (ϕ).

Types of Modulation

Since three characteristics of the sine wave carrier can be varied, it follows that there are three types of modulation. These are amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). However, in practice, it is very difficult to distinguish between phase and frequency modulation. Therefore, these two types of modulation are grouped together under the title of Angle modulation. Thus, there are two basic types of modulation: Amplitude and Angle. The next section will discuss both of these in detail.

SUMMARY

In this chapter you reviewed the characteristics of information and how information content of a message decreases as your ability to predict the remaining part of a message increases.

Because the human language is so redundant, you can miss an occasional word of a conversation and not miss much information. In contrast, computers use digital data which is not redundant. Loss of a single bit of information can completely change the information being conveyed.

Next you learned how the bandwidth of a communications channel affects the speed, and therefore amount, of information that can be sent over the channel. The characteristics of a channel determines how much information can be sent over it.

Telegraph, telephone, and radio increased both the distance and speed at which communications could be transferred from one place to another. Telegraph and telephone both require a pair of wires to work. Radio, however, is a wireless system that works by adding modulation to a carrier wave. The two basic types of modulation are amplitude modulation and angle modulation.

SELF-TEST REVIEW

1. Define communication.

   _____

   _____

2. Define information.

   _____

   _____

3. List at least two reasons why audio information isn't transmitted at its original frequency. _____

   _____

4. What is modulation?

   _____

   _____

5. What is a carrier?

   _____

   _____

6. What three characteristics of a sine wave can be varied or modulated? _____

   _____

   _____

7. List the two basic types of modulation.

   _____

8. A signal with high _____ has a low _____ content.

CHAPTER 2 — Review of Complex Waves

INTRODUCTION

In the first chapter, the need for modulation and its basic principles were established. Now, before we continue our study of modulation, should review complex waves and frequency domain analysis.

Frequency domain analysis stems from the fact that we can think of any periodic waveform as being made up of sine waves. A periodic wave has the same waveshape from one cycle to the next. The period may be very long or very short. All the waveforms shown in Figure 2-1 are periodic waves if they have the same shape on each and every cycle. Any waveform of this type can be formed by superimposing a number of sine waves which have certain amplitude, phase, and frequency characteristics. Similarly, regardless of how a particular periodic waveform is formed, it can be proven that it can be represented by a series of sine waves having specific phase, frequency, and amplitude relationships.

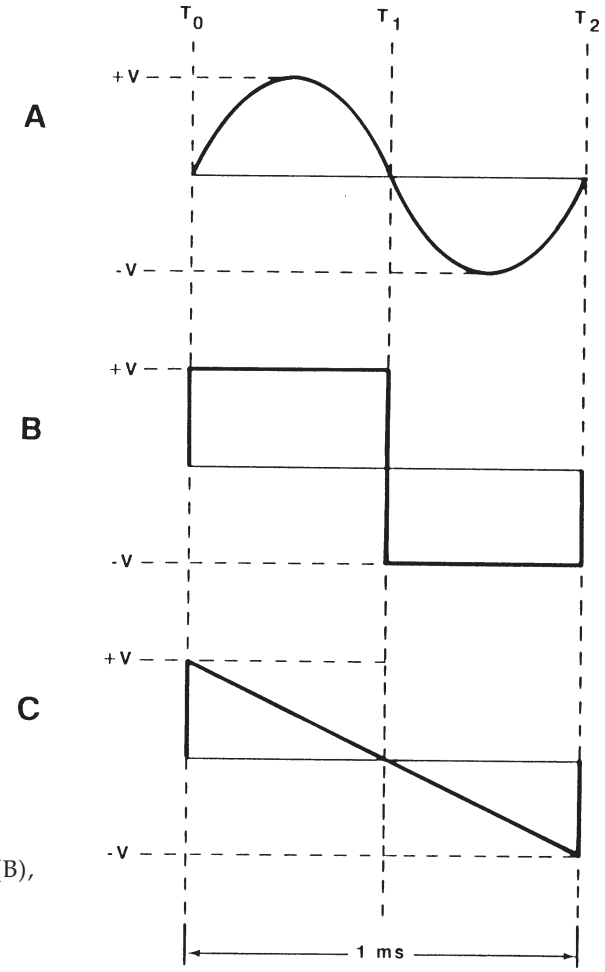


Figure 2-1

A sine wave (A), square wave (B), and sawtooth wave (C).

OBJECTIVES

When you complete this chapter, you will be able to:

- List the characteristics of a sine wave that can be varied.
- State the difference between the time and frequency domains.
- Identify the frequency composition of complex waves including square and sawtooth waves.
- Explain the principle of Fourier analysis.

SINE WAVE

Since the sine wave is the basis of all waveforms, let's examine it in more detail. Figure 2-1A is a time domain representation of the sine wave. It shows the wave as it would appear on an oscilloscope display, which is a graph of time versus amplitude.

Figure 2-2 shows the wave as it would appear on a spectrum analyzer display. This device analyzes the wave's frequency domain. As you can see, it gives us a graph of frequency versus amplitude. It is actually a bar graph showing all the frequency components contained within a wave. Since this sine wave has a period of 1 ms, its frequency is 1 kHz. Therefore, a bar representing the wave's amplitude appears at 1 kHz. There are no other frequency components, since this is a pure sine wave.

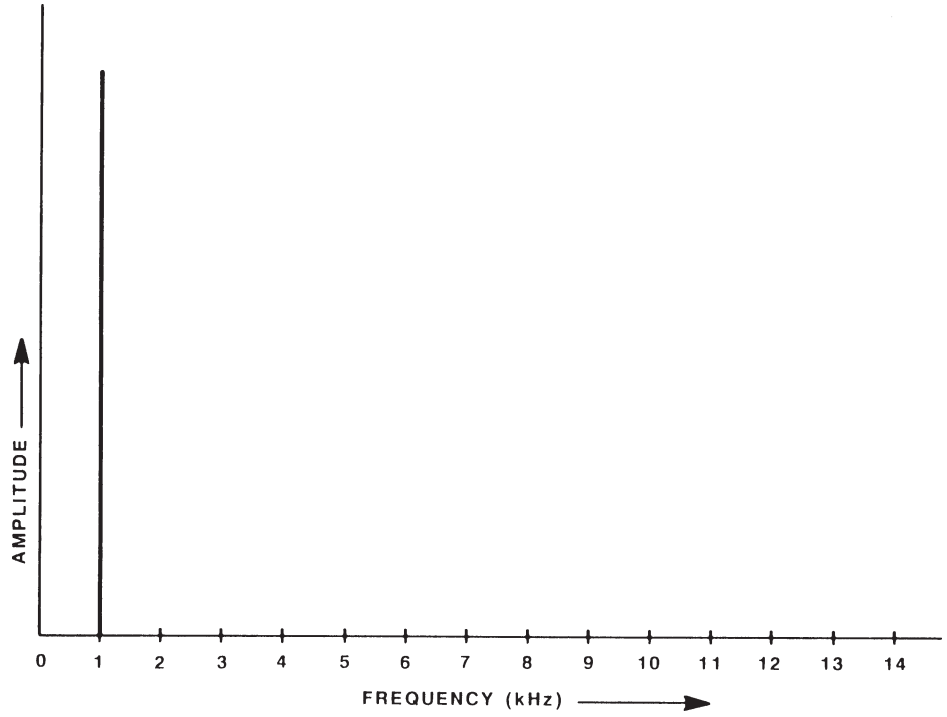


Figure 2-2

Spectrum analyzer display of frequency domain.

SQUARE WAVE

Figure 2-3A shows one cycle of a 1 kHz square wave. Since this is a periodic wave, this cycle is repeated over and over again. This waveform is made up of a large number of sine waves. More specifically, a perfect square wave is composed of a fundamental frequency and an infinite number of odd harmonics.

The result of adding the first harmonic or fundamental and the third harmonic is shown in Figure 2-3B. The addition of the fifth harmonic is shown in Figure 2-3C. Note that the resultant waveform is already approaching a true square wave. In fact, as more odd-order harmonics are added, the closer the waveform gets to becoming a "perfect" square wave.

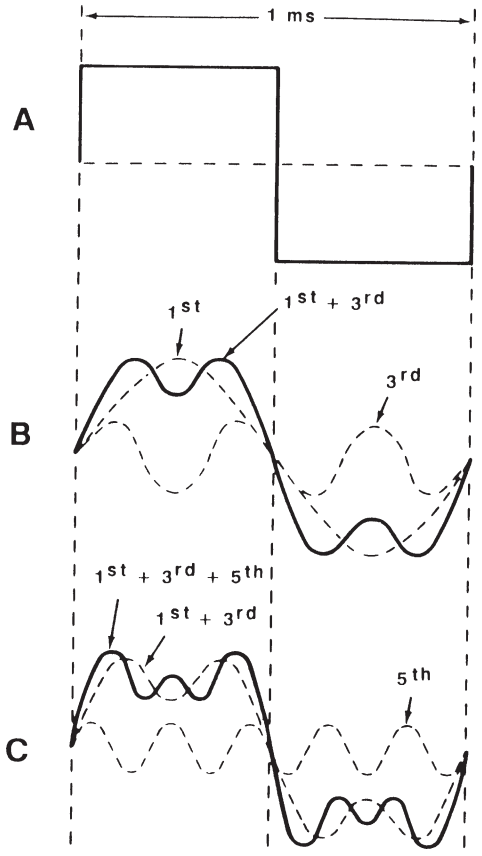


Figure 2-3
The composition of a square wave.

It should now be a simple matter to plot the frequency domain of a square wave since we know that it is composed of odd-order harmonics. Figure 2-4 shows the results. It illustrates the correct amplitude for all the harmonics out to 13 kHz. For a perfect square wave, the chart would extend to infinity.

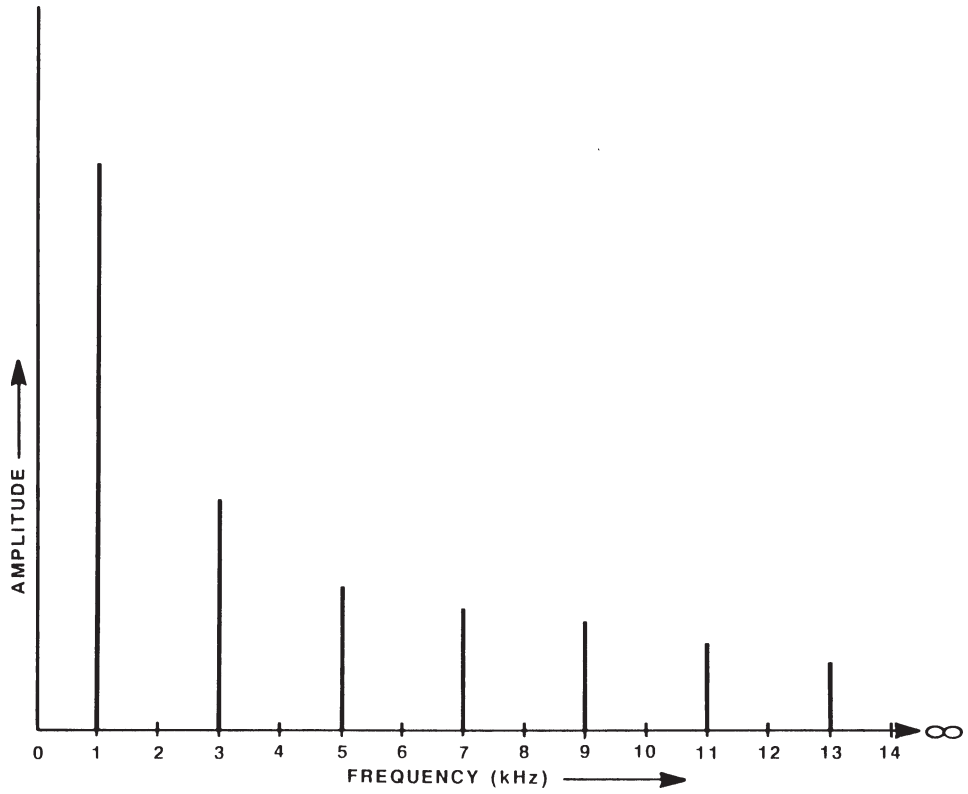


Figure 2-4

The frequency domain analysis of a square wave.

SAWTOOTH WAVE

Figure 2-5 illustrates that a different type of waveform is developed when the harmonic content is changed. Here both even and odd-order harmonics are added. The result is a sawtooth wave. Figure 2-5C shows that, with just the first through the fourth harmonics present, the resultant waveform comes very close to the perfect sawtooth.

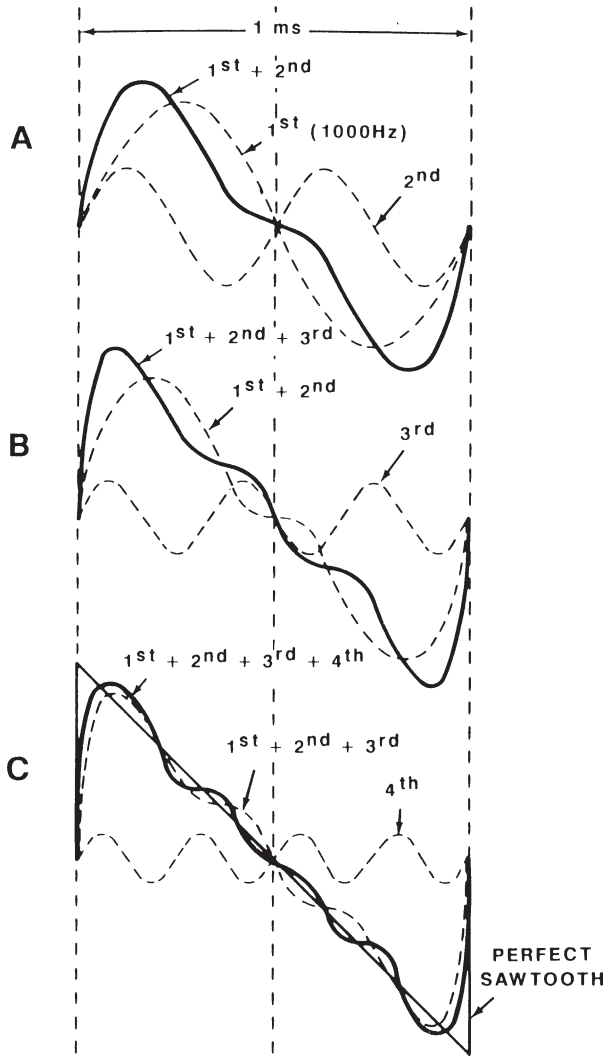


Figure 2-5

The superimposing of sine waves to form a sawtooth wave.

The frequency domain analysis of a sawtooth is shown in Figure 2-6. Note that all harmonics are present, both even and odd-ordered, and that their relative amplitudes differ from the square wave frequency chart.

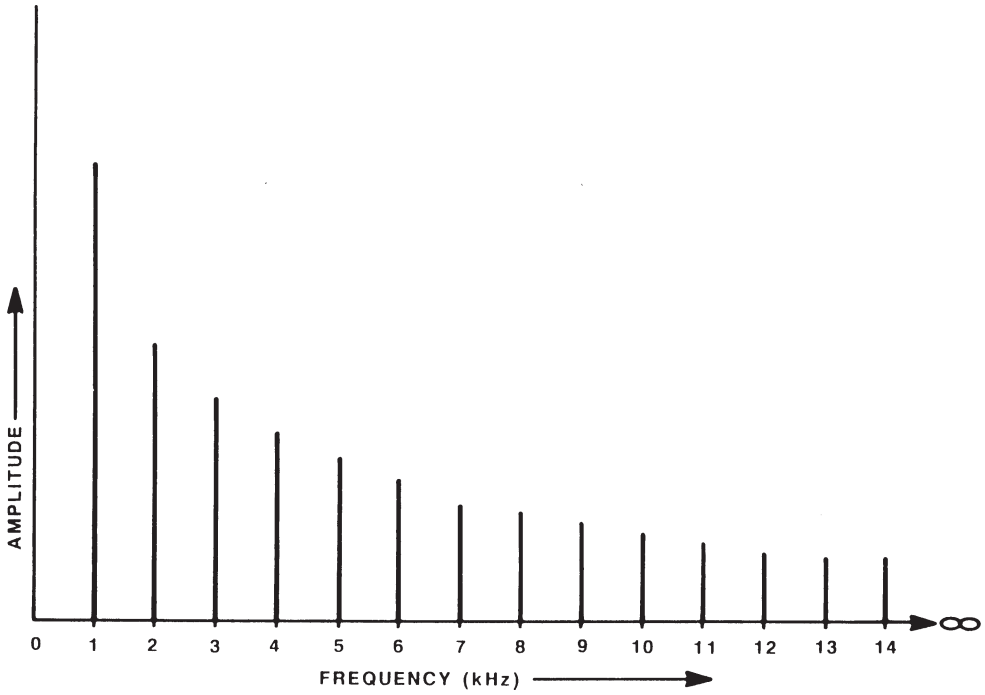


Figure 2-6

The frequency spectrum of a sawtooth wave.

The sawtooth and square wave are only two examples of the unlimited variety of complex waveforms. The important concept is that sine waves can be combined in an infinite number of ways to produce an infinite variety of waveforms. By the same token, any waveform can be broken down into its sine wave components.

FOURIER ANALYSIS

This ability to represent complex waves by a series of sine waves is critical to understanding electronic communication. As it derives from the work of the mathematician Fourier (pronounced foor' yay), it is called Fourier analysis.

Representing complex waves by a series of sine waves is no mathematical trick. If you take signal generators which produce sine waves of the appropriate frequency, phase and amplitude, you can add their outputs together in the laboratory to produce any complex wave you like.

That's right, ANY complex wave. By the simple expedient of assuming the period of the wave to be infinitely long, you can use Fourier analysis to describe a single pulse. This is extremely valuable to communications designers, as only by knowing the frequencies a channel must pass can the channel be designed to handle signals properly.

What bandwidth is needed if a channel is to accurately pass a square wave? Without the tools provided by Fourier analysis, you can only guess. But using Fourier analysis, you can decide in advance what level of fidelity you will accept for your square wave and that immediately determines how many harmonics of the base frequency the channel must pass. Simply and effectively, you have defined the channel bandwidth required for a complex wave. The same procedure works for any complex wave, which is why it is so valuable.

Conversely, if you know the channel characteristics, you can use Fourier analysis to describe the effect of the channel on any particular complex wave that interests you. Since there will always be an infinite number of sine waves needed to describe a complex wave perfectly, and since you cannot ever transmit an infinite number of sine waves, compromises will be necessary. The basic compromise is usually the choice of bandwidth.

BANDWIDTH OF COMPLEX SIGNALS

As you have seen, we generally define bandwidth as the width of a signal between the 3 dB points of the waveform. That is, between the frequencies where the power of the signal has fallen to one-half of its maximum value. If you can capture the signal between the 3 dB points, you have gotten enough of it to accurately interpret its contents.

With complex signals, things are not quite so straightforward. Where are the 3 dB points of a square wave? You saw that the more harmonics we transmit, the better the representation of the pulses at the receiver. However, the transmission path has no idea of what sort of signal you intend to send over it. Nor does it much care. The transmission path is an unthinking channel of a given bandwidth. Anything outside that bandwidth is not going to get through. So how much bandwidth is enough for complex waves?

As a practical matter, we can say that the bandwidth of a complex signal is that bandwidth that contains enough of the sine wave components so that we can reconstruct the complex signal accurately enough for our purposes. For example, all that is necessary for the proper detection of data pulses is to dependably detect their existence. Once we know they are there, we can process the detected signal to get nice, sharp-edged pulses for use elsewhere in the system.

There is, therefore, no hard and fast answer to the question "How many harmonics are enough?" for a complex wave. You have seen from Figure 2-3B that with as few as two harmonics (the first and third) you can probably detect a square wave pulse train accurately, and you can almost certainly do so if you include the fifth harmonic. Similarly, the first three or four harmonics provide a workable representation of a sawtooth wave. Although much depends on the specifics of the system, a useful rule of thumb is that the bandwidth of a pulse train is from 3 to 5 times the pulse repetition frequency.

No matter what form a signal takes, knowing the bandwidth required to accurately transmit it is critical in communication systems. Fourier analysis allows you to determine that bandwidth quickly and accurately.

SUMMARY

In this chapter we reviewed complex waves and the difference between time and frequency domains. You learned that any periodic signal can be created by adding together the correct number of sine waves in the proper phase and amplitude relationship.

You learned that Fourier analysis helps us determine the bandwidth that is required for a channel to accurately pass complex waves. It also allows us to determine how the characteristics of a channel affect a particular complex wave.

SELF-TEST REVIEW

1. What is the difference between the time and frequency domains?

   _____

   _____

2. Any complex waveform is composed of _____

   _____

3. A square wave is composed of _____

   _____

4. A sawtooth wave is composed of _____

   _____

5. What is the bandwidth of a square wave? _____

   _____

CHAPTER 3 — Amplitude Modulation

INTRODUCTION

As we stated earlier, with amplitude modulation the carrier's amplitude is varied in accordance with the modulating signal. There are several circuits that accomplish this and they will be examined in the next unit. Right now, we wish to limit our discussion to the characteristics of the modulated waveform itself.

OBJECTIVES

When you complete this chapter, you will be able to:

- Measure the percent of modulation when given an AM waveform.
- Find the frequencies of AM sidebands when given the carrier and modulating frequencies.
- Find the bandwidth of an AM signal.

THE AM WAVEFORM

Figure 3-1 shows a very simple AM circuit. Here, a radio frequency carrier is applied at "A" and the modulating audio tone at "B". The circuit consists of a nonlinear device such as a diode or transistor. The two signals "mix" in this circuit and produce the AM waveform shown at "C". Notice that both the negative and positive peaks of the output waveform correspond exactly to the modulating tone's waveform.

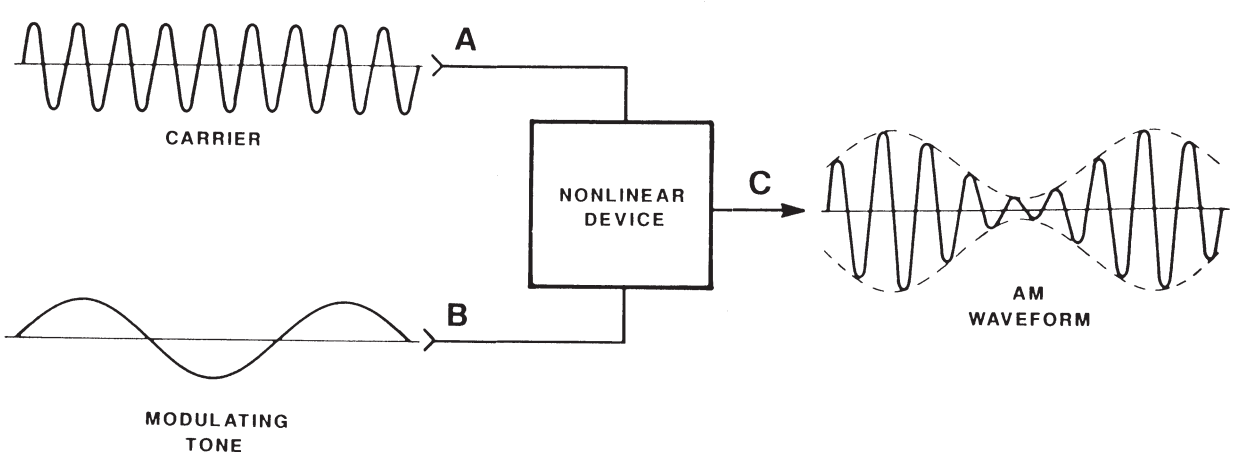


Figure 3-1

The basic method or obtaining amplitude modulation.

The amplitude and frequency of the modulating tone determines the shape of the output waveform or the modulation envelope. For example, Figure 3-2A shows a high amplitude audio signal. The resultant modulated waveform is shown in Figure 3-2B. On the other hand, Figure 3-2C shows a low amplitude, higher frequency audio signal. The modulated waveform is in Figure 3-2D.

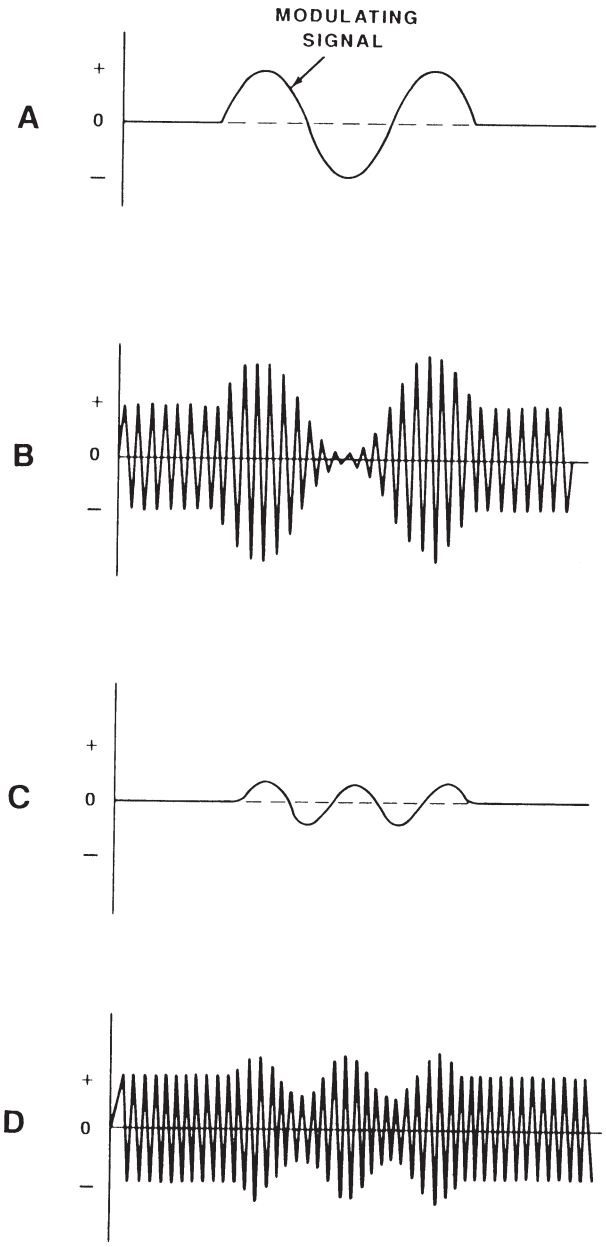


Figure 3-2

Examples of how a modulated waveform varies with the modulating signal.

PERCENT OF MODULATION

The waveforms of Figures 3-2B and D are said to have different degrees of modulation. The degree of modulation is normally expressed as a percentage from 0 to 100. However, it is also known as the modulation factor which varies from 0 to 1. An unmodulated carrier like that shown in Figure 3-3A has 0% modulation. For comparison purposes, let's assume that the carrier has a peak-to-peak amplitude of 40 volts as shown.

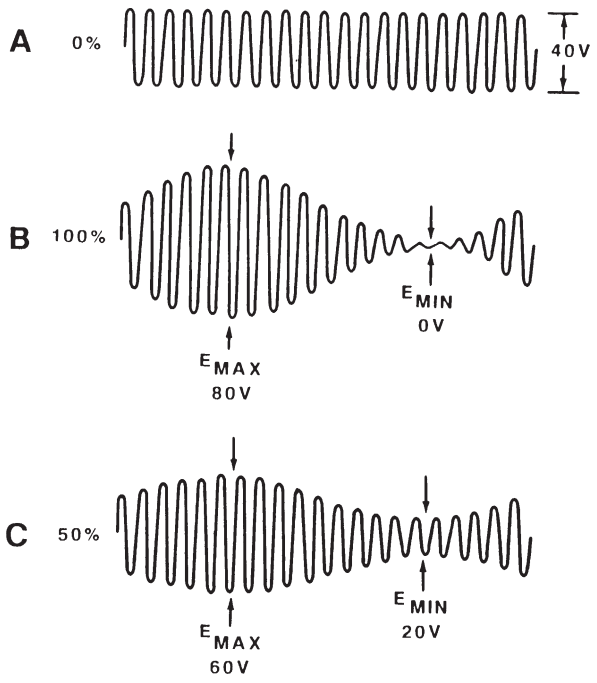


Figure 3-3

Measuring the percent of modulation.

Figure 3-3B shows the same carrier modulated to 100%. Here, the amplitude of the modulated waveform falls to zero volts for an instant during each cycle of the modulating wave. Also, the amplitude increases to 80 volts peak-to-peak once during each cycle of the modulating wave. The average peak-to-peak amplitude is still 40 volts.

In Figure 3-3C the carrier is shown modulated to 50%. The peak-to-peak amplitude varies from 60 volts to 20 volts. However, the average peak-to-peak amplitude is still 40 volts.

The equation for determining the percent of modulation is

$$\text{Percent of modulation} = \frac{E_{max} - E_{min}}{E_{max} + E_{min}} \times 100$$

For example, in Figure 3-3C

$$\%M = \frac{E_{max} - E_{min}}{E_{max} + E_{min}} \times 100$$

$$\%M = \frac{60\ V - 20\ V}{60\ V + 20\ V} \times 100$$

$$\%M = \frac{40\ V}{80\ V} \times 100$$

$$\%M = 0.5 \times 100 = 50$$

Generally, it is desirable to keep the percent of modulation high. For a given transmitter power, a high percent of modulation produces a stronger audio tone in the receiver. The reason for this can be visualized from Figure 3-4. Since the AM receiver recovers just the modulation envelope of the transmitted wave, it is easy to see that the higher modulated waveform in "B" will produce a louder signal than that in "A".

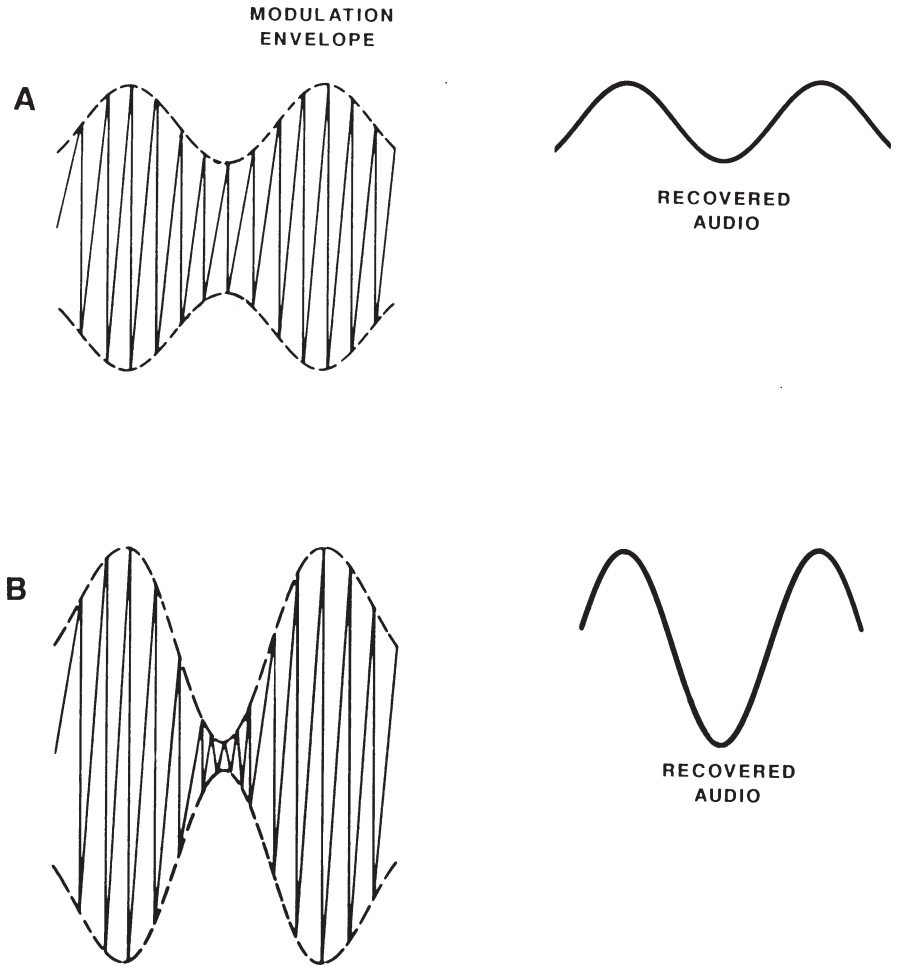


Figure 3-4

The relative amplitude of the recovered audio depending upon the modulation percentage.

While it is a good idea to keep the percent of modulation high, over modulation must be avoided. Overmodulation is shown in Figure 3-5C. It occurs when the amplitude of the modulating signal is too high compared to the unmodulated carrier. Obviously, the minimum amplitude of the carrier is zero volts. It cannot drop below this level regardless of how high the modulating signal is. If the modulating signal is too high, it causes the carrier to cut off for a portion of each cycle. As a result, part of the envelope is distorted. That is, the envelope is not an accurate representation of the modulating wave.

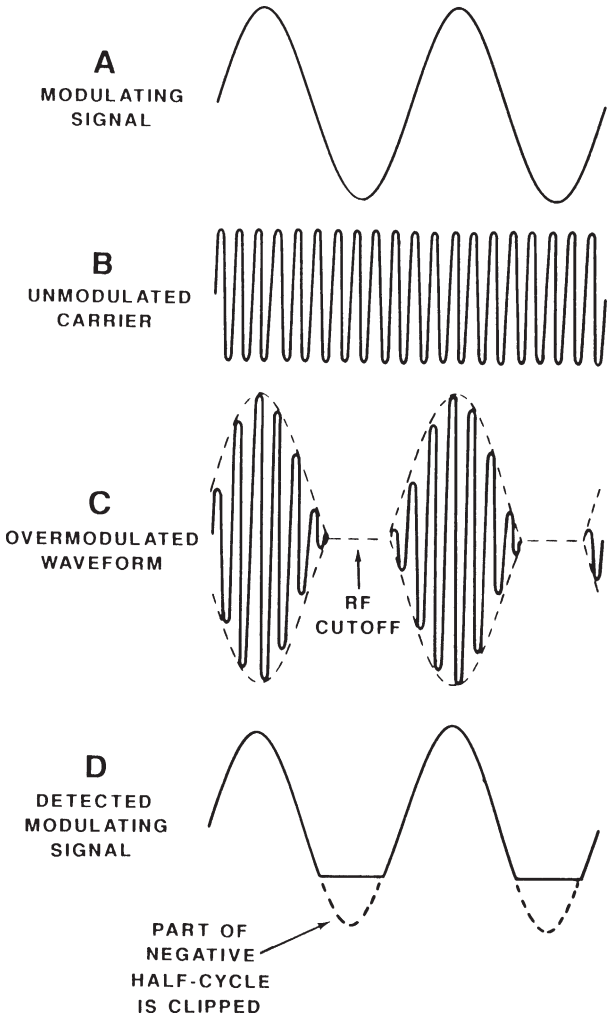


Figure 3-5

Overmodulation causes severe distortion in the received signal.

Figure 3-5A shows the high modulating waveform. The unmodulated carrier is shown in Figure 3-5B. The modulated waveform, shown in Figure 3-5C, cuts off for a portion of each cycle. At the receiver, the envelope is detected, and since it is distorted, the detected waveform is also distorted. The detected envelope is shown in Figure 3-5D.

SIDEBANDS

In Chapter 1 you learned that any complex waveform can be broken down into its component sine waves. The same is true for an amplitude modulated waveform such as that shown in Figure 3-6A. This wave is a 1 MHz carrier modulated by a 10 kHz sine wave. At first glance, you might say that the modulated wave is composed of a 1 MHz sine wave and a 10 kHz sine wave. However, if we apply the waveform to both a 10 kHz bandpass filter and a 1 MHz bandpass filter as shown in Figure 3-6B, we see an output only at the 1 MHz filter. This shows that there is no 10 kHz signal present in the modulated wave; only the modulation envelope represents the audio signal. But we know that some other sine wave components must exist in this complex wave.

If we had a tunable bandpass filter, or a spectrum analyzer, we could search the spectrum and determine what other frequencies are contained within the modulated signal. Doing this, we would find a signal at 1.01 MHz and another signal at 0.99 MHz. These two signals are called sidebands.

> The signals that appear above and below the carrier that occur when the carrier is modulated are called *sidebands*.

They can be extracted from the modulated waveform by using sharply tuned filters as shown in Figure 3-6C.

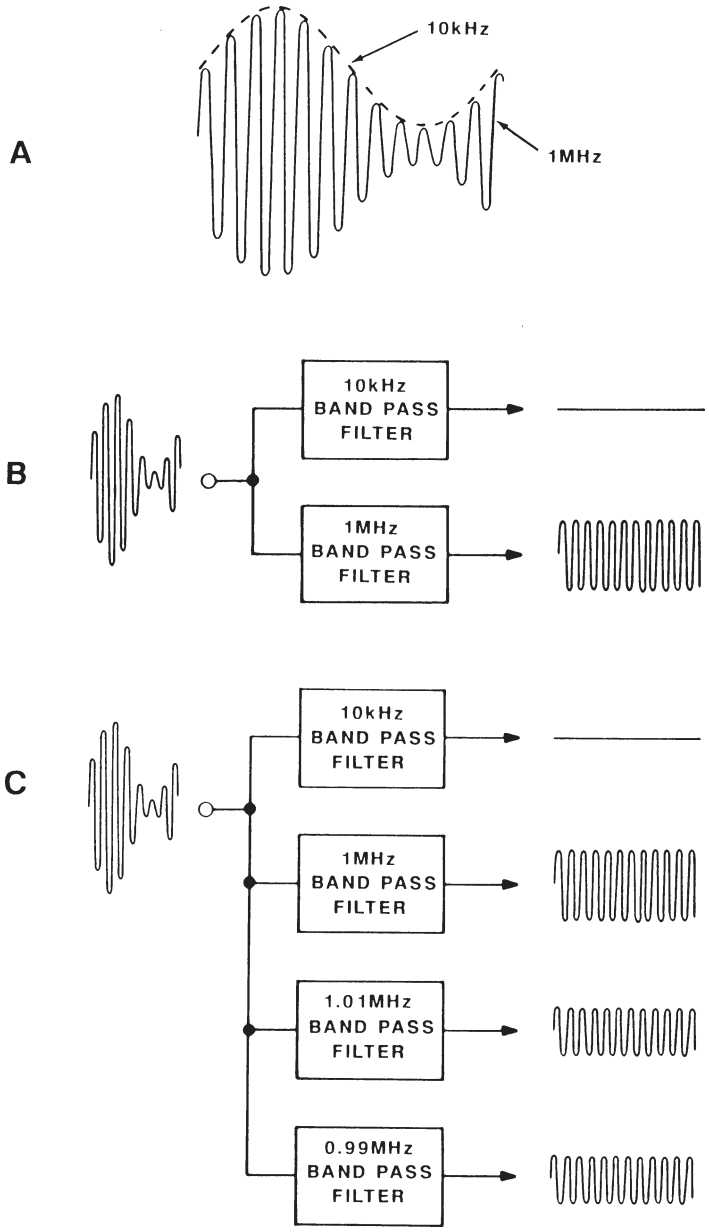


Figure 3-6

Frequency domain analysis of an AM wave.

The higher frequency (1.01 MHz) is called the upper sideband. Its frequency is always equal to the carrier frequency plus the modulating frequency. That is:

$$\text{Upper sideband} = f_c + f_m$$

Where:

f_c = carrier frequency

f_m = modulating frequency

In our example:

$$\text{Upper sideband} = 1\ \text{MHz} + 10\ \text{kHz}$$

$$= 1.01\ \text{MHz}$$

The lower frequency (0.99 MHz) is called the lower sideband. Its frequency is equal to the carrier frequency minus the modulating frequency. In this case, it is:

$$\text{Lower sideband} = f_c - f_m$$

$$= 1\ \text{MHz} - 10\ \text{kHz}$$

$$= 0.99\ \text{MHz}$$

If we were to carefully observe the filter outputs of Figure 3-6C, we would find that the sideband and carrier amplitudes do not vary. In fact, we would find that the carrier amplitude never varies whether it is modulated or not. You might ask, if the individual frequency components do not change, how can the modulated waveform change to follow the modulating signal? Figure 3-7 gives a detailed look at exactly what happens. It shows that the constant amplitude sidebands are at different frequencies and, therefore, are in phase and out of phase with one another at various times. At time A, the constant amplitude carrier is also in phase. The result is a high amplitude peak in the modulated waveform. Now observe the wave relationships at point B. Once again, the sidebands are in phase with each other. However, they are 180° out of phase with the carrier wave. The result is a low point, or trough, in the modulated waveform.

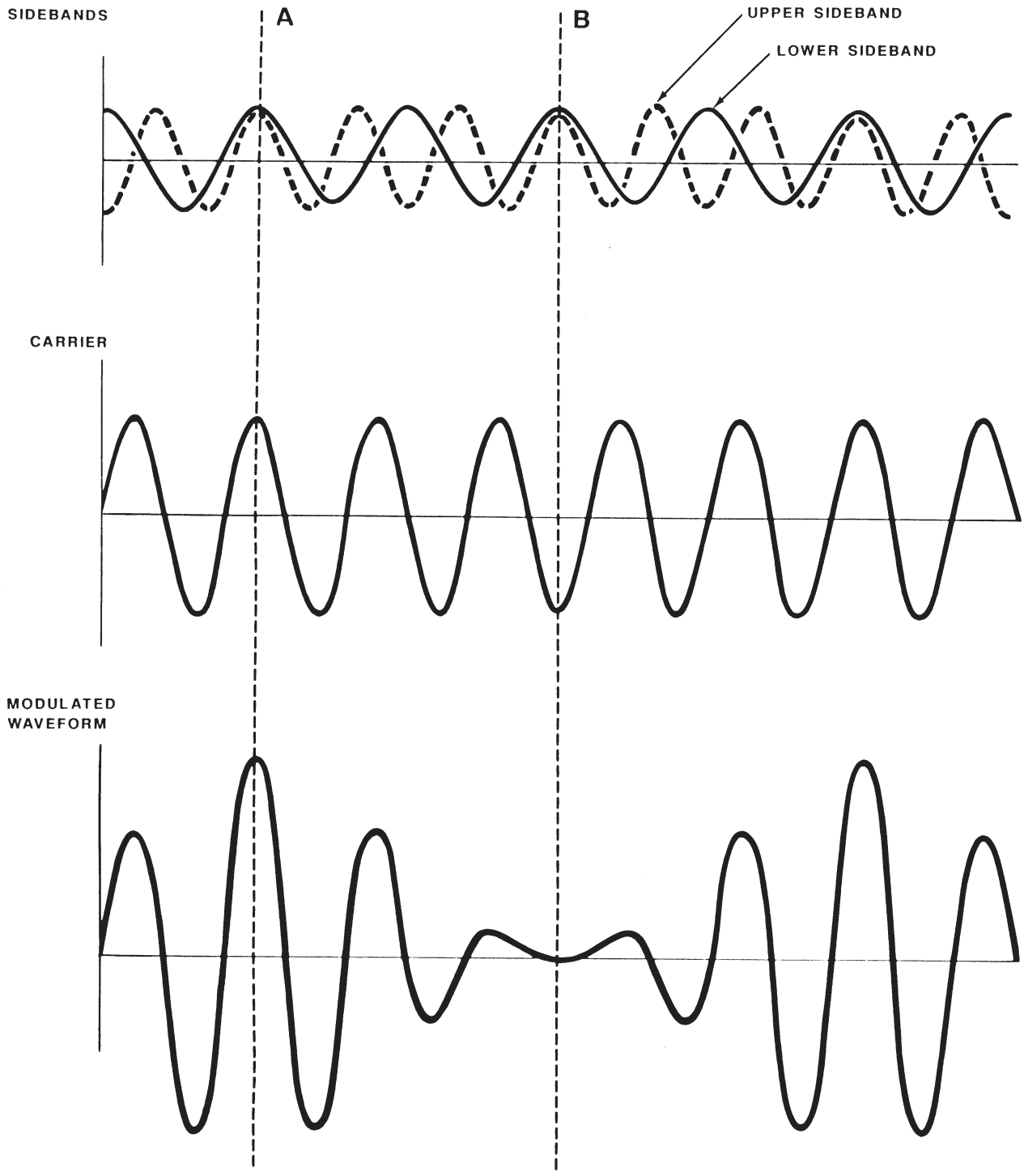


Figure 3-7

The phase relationships of an AM wave.

From this analysis, you can see that the shape of the modulation envelope is dependent on the sidebands. And, the sidebands are in turn dependent upon the modulating signal. That is, the frequency of the sidebands determines their phase relationship and, therefore, the peaks and troughs, or frequency, of the modulation envelope. The sideband amplitude also determines the envelope's amplitude, or percent of modulation. This is because they are either adding to or subtracting from the constant amplitude carrier. This illustrates an important fact about amplitude modulation: the modulating intelligence or information is contained only in the sidebands.

The frequency spectrum charts of Figure 3-8 further illustrate this point. Since these are voltage diagrams, the sideband amplitudes shown add to or subtract from the carrier to produce the modulated envelope. For example, Figure 3-8A shows the sideband amplitudes as being exactly one half that of the carrier. This is the condition for 100% modulation, because when all signals are in phase, the waveform amplitude is twice the carrier and when the sidebands are out of phase with the carrier, the waveform amplitude is zero.

Figure 3-8B shows a 50% modulated signal. Note that the carrier amplitude remains the same while the sideband amplitudes have decreased. The frequency of the sidebands has also changed. The sidebands are further from the carrier because the modulating frequency has increased. This is shown in the modulated waveform to the right.

The result of a square wave modulating signal is shown in Figure 3-8C. In this case, since the modulating signal is actually the fundamental and all odd-order harmonics, there is a sideband for each sine wave component of the modulating signal.

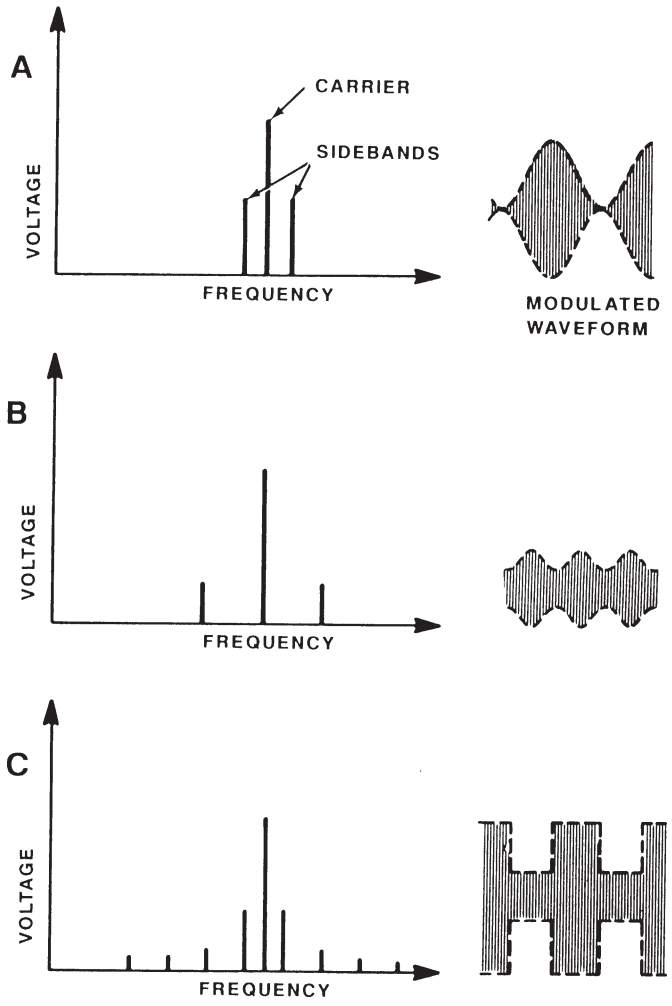


Figure 3-8

The sideband spectrum of AM waves.

BANDWIDTH

You can see from the frequency spectrum charts shown in Figure 3-8 that, with amplitude modulation, the transmitted signal is actually a band of frequencies rather than just the carrier. The carrier contains no information. If we transmitted or received just the carrier, no information would be conveyed.

The bandwidth of an AM signal is determined by its sidebands. It extends from the lowest sideband frequency to the highest sideband frequency. Therefore, the AM signal bandwidth is always twice the highest modulating frequency. If the highest modulating frequency is 15 kHz, then the bandwidth of the modulated signal is 30 kHz.

The width of the sidebands, in turn, depends upon the bandwidth of the modulating signal. In the case of square wave modulation in the preceding paragraph, if that square wave had the same 10 kHz frequency as the sine wave we examined earlier in this section, the square wave modulated carrier would have a much wider bandwidth than the sine wave modulated carrier. That's because it is necessary to transmit sine wave components of 30 kHz, 50 kHz, and possibly higher for the square wave, where only the 10 kHz component is needed to represent the pure sine wave.

In practice, pure sine waves are not used to modulate carriers except as test signals. Signals used to modulate carriers are usually complex waves; the most prevalent signal, human speech, is a very complex mixture of complex waveforms. Thus, it is customary to find that the sidebands themselves comprise a band of many frequency constituents, rather than one or a few easily identified sine wave components. Regardless of how many frequency components actually are found in each sideband, however, there are only two sidebands to an AM signal: upper and lower.

In practical radio systems, it is customary to limit the range of frequencies that can modulate a carrier. If this were not done, the transmitted width of the modulated signal could not be predetermined, and it would likely interfere with other signals sharing the same radio frequency band. Consider a familiar example: AM broadcast stations in North America are assigned frequencies 10 kHz apart (980 kHz, 990 kHz, 1000 kHz, etc.). Therefore, no station can broadcast

a signal with a bandwidth greater than 10 kHz, or it could interfere with the station on the adjoining frequency. So it follows that the maximum modulating frequency cannot be allowed to be greater than 5 kHz.

While limiting the input bandwidth makes sense from the standpoint of minimizing interference between modulated waves, it also limits the fidelity of the transmitted signal. We find that practical considerations in the design of receivers limit the bandwidth obtainable there, too, further reducing the maximum possible fidelity of the received signals, and the maximum rate of information transfer over the channel.

These facts have far-reaching implications for designers, users, and maintainers of electronic communication systems. There is no such thing as a perfect communications system. Design compromises and the characteristics of the devices used in the system place limits on performance. Once you have selected a communications system, you are not at liberty to choose the information rate you require; the system designer has already chosen the upper limit. We will see later in this course that methods exist to partially circumvent these restraints, but all of them involve increased complexity and cost.

SUMMARY

In this chapter you learned about amplitude modulation (AM) and how the amplitude and frequency of the modulating tone affects the shape of the modulation envelope.

You also learned how to use the minimum and maximum voltages of a modulated waveform to determine the percent of modulation. In order to produce a stronger signal in a receiver, you generally want to keep the percentage of modulation as high as possible without causing overmodulation.

Next you learned that an amplitude-modulated carrier produces sidebands at frequencies that are the sum and difference between the carrier frequency and the modulating frequency.

Finally you learned that the bandwidth of an AM signal is determined by its sidebands. The bandwidth of an AM signal is always twice the highest modulating frequency.

SELF-TEST REVIEW

1. Find the percent of modulation for the waveforms shown in Figure 3-9.

   A. _____
   B. _____
   C. _____

2. List the three frequency components of an AM wave. _____

   _____

3. The intelligence in an AM signal is contained in the _____

   _____

4. What is the bandwidth of an AM signal when the highest modulating frequency is 6 kHz? _____

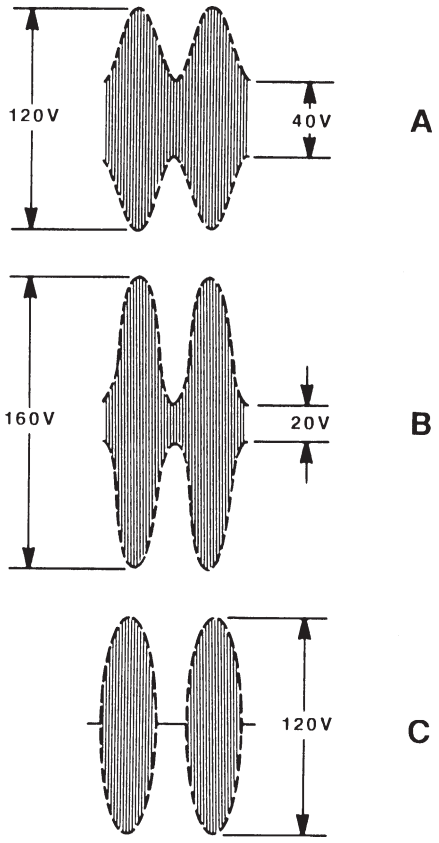


Figure 3-9
What is the percent of modulation for these waveforms?

CHAPTER 4 — Angle Modulation

INTRODUCTION

There are three characteristics of a sine wave that can be modulated: amplitude, angular velocity or frequency, and phase angle. The last two are similar in that the angle velocity or phase is being modulated. They are also similar in that the modulation of one indirectly causes the other to vary. Therefore, they are grouped together under the heading of Angle Modulation.

OBJECTIVES

When you complete this chapter, you will be able to:

- Define angle modulation.
- List the characteristics of frequency modulation.
- State the differences between phase and frequency modulation.
- Find the modulation index, sideband distribution, and bandwidth of an angle-modulated signal.

FREQUENCY MODULATION

Figure 4-1 shows a frequency modulated or FM waveform. The information or modulating waveform is shown in Figure 4-1A, while the unmodulated carrier is shown in Figure 4-1B. With FM, the modulating signal changes the frequency of the carrier rather than its amplitude. The resulting frequency modulated waveform is shown in Figure 4-1C.

At time T_0 the modulated waveform is at its center frequency. As the modulating signal swings positive, the frequency of the carrier increases. The carrier reaches its maximum frequency when the modulating signal reaches its maximum amplitude, at time T_1.

At time T_2, the modulating signal returns to 0 and the carrier returns to its center frequency. After T_2, the modulating signal swings negative. This forces the carrier below its center frequency. The carrier again returns to its center frequency when the modulating signal returns to 0 volts at time T_4. Between times T_4 and T_8, the modulating signal repeats its cycle. As a result, the carrier again shifts in frequency. It swings first above then below its center frequency. Notice that it returns to its center frequency each time the modulating signal passes through 0 volts.

The carrier changes equally above and below its center frequency. The amount of frequency change is called the frequency deviation. For example, let's assume that a carrier continuously swings from 100 MHz, down to 99.9 MHz, back to 100 MHz, up to 100.1 MHz, and back to 100 MHz. The frequency deviation is 0.1 MHz or 100 kHz.

> The amount a carrier changes above and below its center frequency is called frequency deviation.

The rate of frequency deviation is determined by the frequency of the modulating signal. For example, if the modulating signal is a 1 kHz audio tone, the carrier swings above and below its center frequency 1000 times each second. A 10 kHz audio tone still causes the carrier to deviate 100 kHz; but this time at the rate of 10,000 times each second. Thus, the frequency of the modulating signal determines the rate of frequency deviation but not the amount of deviation.

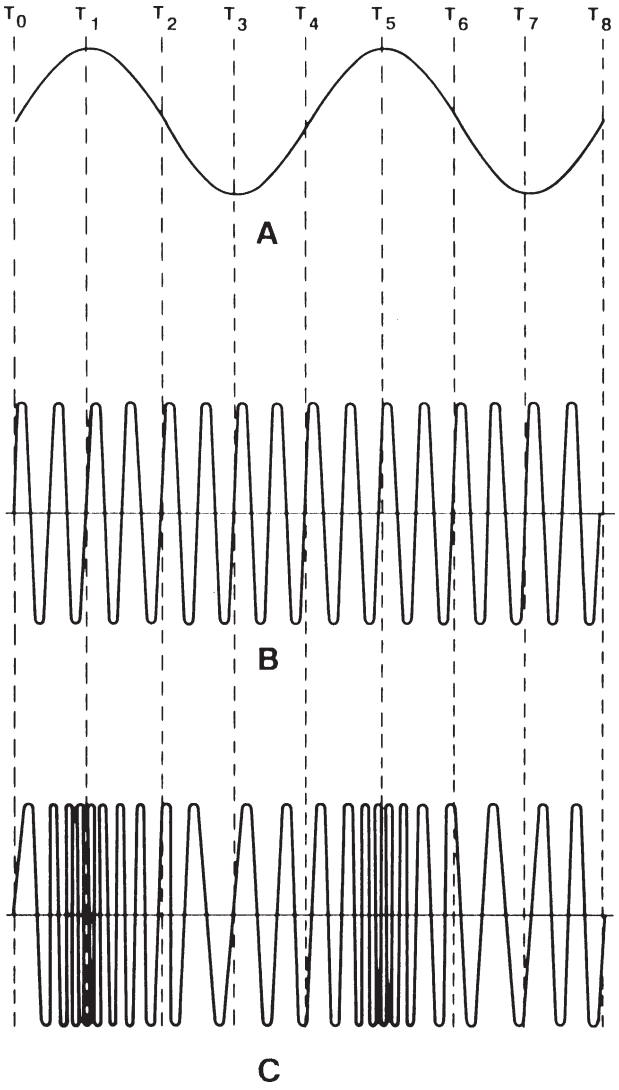


Figure 4-1

Modulating signal (A), the unmodulated carrier (B), and the frequency-modulated waveform (C).

The amount that the carrier deviates from its center frequency is determined by the amplitude of the modulating signal. A high amplitude audio tone may cause a deviation of 100 kHz. A lower amplitude tone of the same frequency may cause a deviation of only 50 kHz.

Thus, the frequency-modulated waveform has the following characteristics:

A. It is constant in amplitude but varies in frequency.

B. The rate of carrier deviation is the same as the frequency of the modulating signal.

C. The amount of carrier deviation is directly proportional to the amplitude of the modulating signal.

Because the information in an FM signal is contained in its instantaneous frequency, and not its amplitude, FM is much more immune to noise than AM. That is why you hear less static on FM radio stations than on AM stations.

PHASE MODULATION

In phase modulation (PM), the carrier's instantaneous phase angle is made to vary in proportion to the modulating signal's amplitude. However, an instantaneous phase change cannot occur without a corresponding frequency change. Therefore, phase modulation also produces carrier frequency deviation. In fact, frequency modulation cannot occur without a certain amount of phase variation. Thus, any variation in frequency or phase produces both FM and PM; the two must occur simultaneously.

A comparison of phase and frequency modulation is shown in Figure 4-2. Careful analysis of the two modulated waveforms shows them to be essentially identical. The difference is that, with phase modulation, the carrier frequency deviation is proportional to the modulation signal's phase change as well as its amplitude. For this reason, in phase modulation, maximum deviation occurs when the modulating signal crosses the zero axis. In other words, the greatest phase change occurs when the modulating wave changes polarity from + to – and from – to +.

This is shown in Figure 4-2 when, at time T_2, the modulating wave crosses the zero axis. The result is that the PM wave is deviated to its lowest frequency. At time T_4, when the modulating signal changes from – to +, the PM wave is deviated to its maximum frequency. Thus, the modulating signal, while phase modulating the carrier, also indirectly frequency modulates it. For this reason, phase modulation is sometimes referred to as indirect FM.

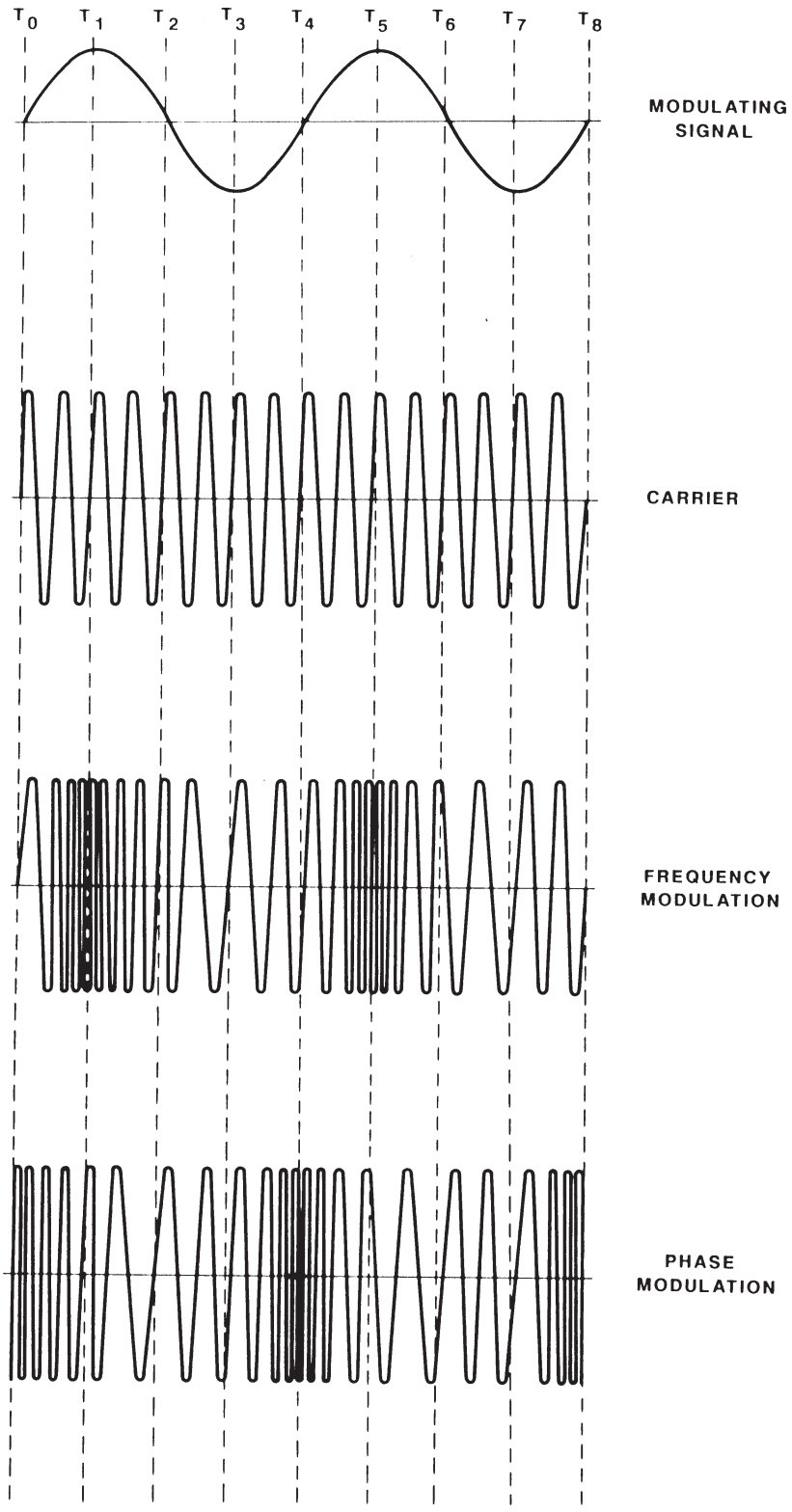


Figure 4-2

The differences between phase and frequency modulation.

As stated earlier, with phase modulation the amount of carrier deviation is proportional to the modulating signal's phase change. This is why maximum deviation occurs when the modulating wave crosses the zero axis. However, consider what happens when the modulating frequency increases. For example, what happens if the modulating frequency increases from 100 Hz to 1000 Hz? At 100 Hz, the signal's phase changes 100 Hz × 360° or 36,000 degrees per second. At 1000 Hz, the phase change increases by a factor of 10 to 360,000 degrees per second. Since the carrier's deviation is proportional to phase change, its deviation increases as the modulating frequency increases.

Thus, the primary difference between PM and FM is:

A. With PM, the carrier frequency deviation is proportional to both the modulation frequency and amplitude.

B. With FM, the frequency deviation is proportional to the modulation amplitude regardless of the signal's frequency.

Actually, at a single modulating frequency with the same deviation, FM and PM are indistinguishable. The only difference is that as the modulation frequency increases, the PM waveform's deviation increases. With this in mind, we can design a simple RC network to convert from PM to FM or vice versa.

Figure 4-3A shows this concept in practice. Here, a low pass network attenuates the modulating signal as its frequency increases. The resulting output is then applied to a phase modulator. (The actual modulator circuits are discussed in a later chapter). The output is an FM wave. This is because, as the modulating frequency increases, the low pass network correspondingly reduces its amplitude. The end result is that, as the deviation increases, due to an increase in the modulating frequency, it also decreases due to the lower modulating signal amplitude. The two changes offset one another. Therefore, the output deviation remains constant regardless of modulating frequency. Thus, the output is actually an FM signal.

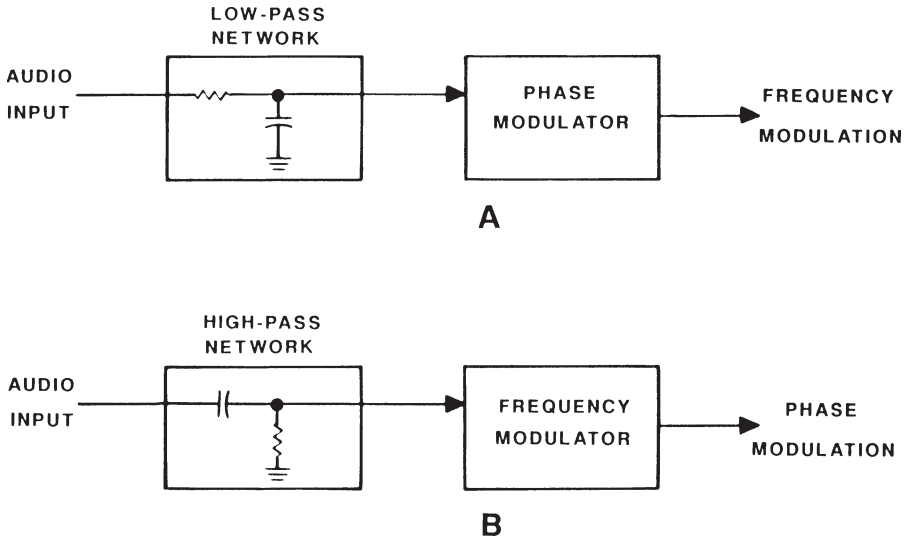


Figure 4-3
Conversion between PM and FM.

The reverse is also possible. Figure 4-3B shows a high pass network used to obtain phase modulation from a frequency modulator. The high pass filter attenuates the low frequencies and allows the high frequencies to pass. Therefore, the output wave's deviation increases as the modulating frequency increases. The output, thus, has the characteristics of phase modulation.

Phase modulation is being increasingly used in communications systems, especially in data modulators and demodulators. The primary reason for this is that it can use a smaller bandwidth to transmit the same information than FM does. This is an attractive property when bandwidth is limited, as it nearly always is.

On the other hand, PM is not especially well-suited to transmitting voice or music, as it is less capable than AM or FM of handling small variations in modulating signal level or frequency (this is essentially due to the difficulty of detecting small changes in phase, and not to any theoretical problems.) As a result, it is almost never used in broadcasting.

Because information is contained in the phase of the signal, and not its amplitude, PM, like FM, is relatively immune to noise. Unlike FM, detecting PM requires that a reference signal phase be determined, otherwise we know nothing about what is changed.

A reference signal can be obtained in one of two ways. First, get the phase from the carrier. Second, establish a local reference at the receiver by measuring the changes in the signal phase. Each of these approaches has advantages and disadvantages, as we will see later in the course.

It is popular to regard phase modulation as merely a special case of FM, and to ignore it. However, the unique characteristics of PM have made it attractive to designers of certain kinds of communication systems, particularly those that handle digital data, and you will see increasing use of PM in such devices as modulator/demodulators.

MODULATION INDEX

In AM, the degree of modulation is measured as a percentage from 0% to 100% or as a modulation factor from 0 to 1. In angle modulation, the degree of modulation is measured by the modulation index. The equation for modulation index is:

$$m = \frac{f_d}{f_m}$$

Where

f_d = The frequency deviation

f_m = The modulating frequency

While the modulation factor in AM is limited to a decimal between 0 and 1, it must be emphasized that the modulation index in angle modulation can reach quite high numerical values. For example, the maximum deviation allowed in FM broadcasting is 75 kHz. If a 1 kHz audio signal causes full deviation, the modulation index is:

$$m = \frac{75\ \text{kHz}}{1\ \text{kHz}}$$

$$m = 75$$

Another measure of angle modulation is the deviation ratio. This is the ratio of the maximum deviation to the maximum audio frequency, thus, it is a total system measurement rather than the instantaneous measurement of modulation index. Using the FM broadcast system as an example, maximum deviation is 75 kHz and the maximum audio frequency is 15 kHz. The deviation ratio is:

$$\text{deviation ratio} = \frac{f_{d\,max}}{f_{m\,max}}$$

$$\text{deviation ratio} = \frac{75\ \text{kHz}}{15\ \text{kHz}}$$

$$\text{deviation ratio} = 5$$

For phase modulation, the modulation index is equivalent to the maximum change in phase. Because we can't distinguish a change of +180 degrees in phase from a change of –180 degrees, the maximum range of phase variation is 180 degrees. Since we can shift phase either plus or minus that amount from the reference phase, the maximum possible phase deviation (and therefore the maximum possible modulation index) is 2×180, or 360 degrees.

Because the PM modulation index cannot be made arbitrarily large, PM cannot be used to spread information over as wide a bandwidth as FM can. Depending upon the application, this can be either an advantage or a disadvantage.

SIDEBANDS

One of the distinct differences between amplitude and angle modulation is the number of sidebands. AM has only two, the upper and lower sidebands. In angle modulation, the number of sidebands is theoretically infinite. This is because the numerous frequency shifts produced by the modulating signal cause the generation of many additional frequencies. Fortunately, many higher order sidebands contain an insignificant amount of energy and can, therefore, be disregarded.

In AM, the sidebands add to or subtract from the constant amplitude carrier which results in the modulation envelope. However, in angle modulation, the waveform remains at a constant amplitude. This means that, as sideband number, amplitude, or distribution change, the carrier must also change to keep the resultant waveform's amplitude constant. This interrelationship between carrier and sidebands is orchestrated by the modulation index. That is, the modulation index determines the number of significant sidebands, their amplitude, and the carrier's amplitude.

A graph of modulation index versus carrier and sideband amplitude is shown in Figure 4-4. You can see that, with a modulation index of zero, the carrier amplitude is one and there are no sidebands. As the modulation index increases, more sidebands are added, their amplitude increases, and the carrier decreases. This happens until, at an index of 2.4, the carrier disappears entirely. This is known as the first carrier "null", since it also occurs at still higher modulation indices. At an index of 3.1 the carrier's amplitude is -0.3. This indicates that it is 180° out of phase with the components above the zero axis.

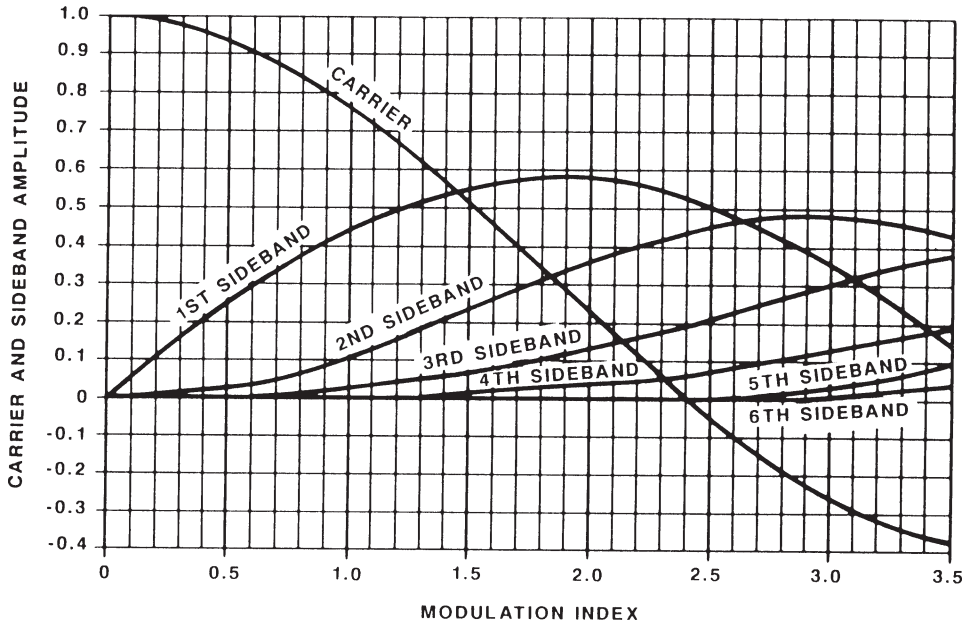


Figure 4-4

Graph of sideband and carrier amplitude in an angle modulated signal.

Figure 4-5 is a table of sideband and carrier amplitudes at various modulation indices. Although the sidebands theoretically stretch out to infinity, any sidebands with an amplitude less than 1% of the original carrier are insignificant and, therefore, left out. As an example, suppose the modulation index is 0.5. From the table in Figure 4-5, a modulation index of 0.5 means that carrier amplitude is 0.94, the first sidebands have an amplitude of 0.24, and the second, and last significant sidebands, have an amplitude of 0.03. This sideband distribution is shown in Figure 4-6, along with several others.

MODULATION INDEX	CARRIER	SIDEBANDS															
		1st PAIR	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th	14th	15th	16th
0.00	1.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.25	0.98	0.12	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.5	0.94	0.24	0.03	—	—	—	—	—	—	—	—	—	—	—	—	—	—
1.0	0.77	0.44	0.11	0.02	—	—	—	—	—	—	—	—	—	—	—	—	—
1.5	0.51	0.56	0.23	0.06	0.01	—	—	—	—	—	—	—	—	—	—	—	—
2.0	0.22	0.58	0.35	0.13	0.03	—	—	—	—	—	—	—	—	—	—	—	—
2.5	-0.05	0.50	0.45	0.22	0.07	0.02	—	—	—	—	—	—	—	—	—	—	—
3.0	-0.26	0.34	0.49	0.31	0.13	0.04	0.01	—	—	—	—	—	—	—	—	—	—
4.0	-0.40	-0.07	0.36	0.43	0.28	0.13	0.05	0.02	—	—	—	—	—	—	—	—	—
5.0	-0.18	-0.33	0.05	0.36	0.39	0.26	0.13	0.05	0.02	—	—	—	—	—	—	—	—
6.0	0.15	-0.28	-0.24	0.11	0.36	0.36	0.25	0.13	0.06	0.02	—	—	—	—	—	—	—
7.0	0.30	0.00	-0.30	-0.17	0.16	0.35	0.34	0.23	0.13	0.06	0.02	—	—	—	—	—	—
8.0	0.17	0.23	-0.11	-0.29	-0.10	0.19	0.34	0.32	0.22	0.13	0.06	0.03	—	—	—	—	—
9.0	-0.09	0.24	0.14	-0.18	-0.27	-0.06	0.20	0.33	0.30	0.21	0.12	0.06	0.03	0.01	—	—	—
10.0	-0.25	0.04	0.25	0.06	-0.22	-0.23	-0.01	0.22	0.31	0.29	0.20	0.12	0.06	0.03	0.01	—	—
12.0	-0.05	-0.22	-0.08	0.20	0.18	-0.07	-0.24	-0.17	0.05	0.23	0.30	0.27	0.20	0.12	0.07	0.03	0.01
15.0	-0.01	0.21	0.04	0.19	-0.12	0.13	0.21	0.03	-0.17	-0.22	-0.09	0.10	0.24	0.28	0.25	0.18	0.12

Figure 4-5

Table of sideband and carrier distribution for several modulation signal.

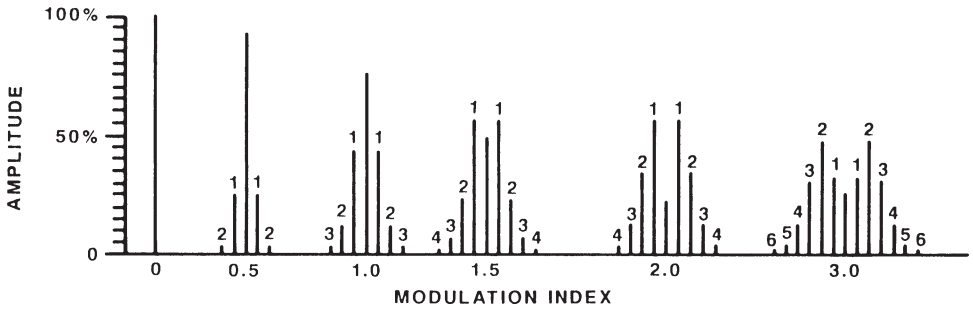


Figure 4-6

Sideband distribution for angle modulation.

BANDWIDTH

In angle modulation, the number of sidebands and their amplitude is determined by the modulation index. However, the frequency of each sideband depends upon the modulating frequency. The first-order sidebands are $f_c + f_{in}$ and $f_c - f_m$. The second-order sidebands are $f_c + 2_{fm}$ and $f_c - 2_{fm}$. This progression continues for each higher order sideband. The bandwidth, therefore, depends upon the number of sidebands in the wave. This is determined by the modulating frequency and modulation index. Since modulation index is $\frac{f_d}{f_m}$, frequency deviation is also a determining factor in bandwidth.

Thus, if we know the modulating frequency and the frequency deviation, we can easily determine the required bandwidth.

As an example, what is the bandwidth of an angle modulated signal in which the modulating frequency is 3 kHz and the maximum deviation is 18 kHz? We must first find the modulation index.

$$m = \frac{f_d}{f_m}$$

$$m = \frac{18\ \text{kHz}}{3\ \text{kHz}}$$

$$m = 6$$

From the table in Figure 4-5, we find that a modulation index of 6, has 9 significant sideband pairs. Therefore, the bandwidth is

$$BW = fm \times \text{highest order sideband} \times 2$$

$$BW = 3\ \text{kHz} \times 9 \times 2$$

$$BW = 54\ \text{kHz}$$

ANGLE MODULATION AND PULSES

It is instructive to look at what types of signals are produced using the three types of modulation (AM, FM, and PM) when the modulating signal is a pulse stream. You can see the time domain comparison in Figure 4-7.

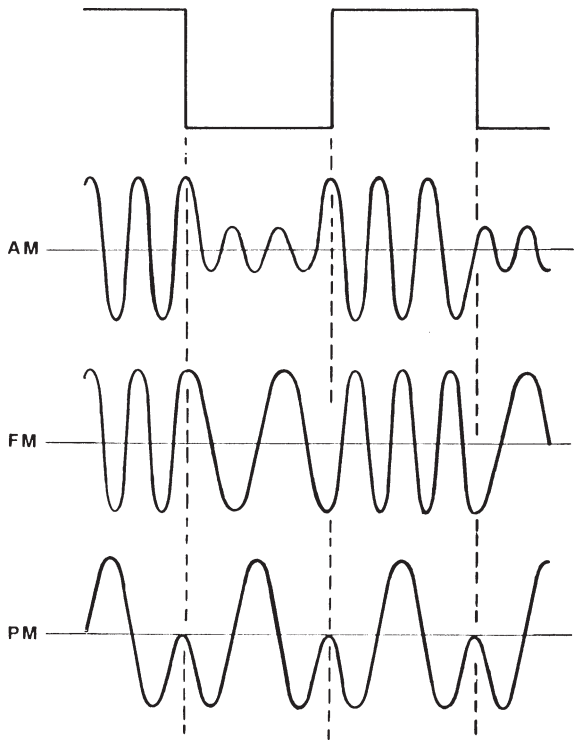


Figure 4-7
Comparison of pulse modulation: AM, FM, & PM.

In Figure 4-7, notice that neither the frequency nor the phase of the AM signal changes at the pulse boundaries, only the envelope amplitude. This is what you would expect. In the case of FM, the carrier frequency changes at the pulse boundaries. You can't see it in this diagram, but the phase also changes slightly; however, the important thing is that the phase of the FM signal is continuous. There is no sudden change in the phase of the carrier, but rather in its frequency.

Quite the opposite is true for PM. There is an abrupt 180-degree carrier phase change at each pulse boundary. Obviously, there is also some frequency change (you can't have one without the other), but the phase change is the predominant effect, and the one that the receiver will be looking for. In the case of pulses, PM's insensitivity to subtle changes in the modulating signal is an advantage: there is nothing subtle about pulse transitions, nor the abrupt phase change they cause in the carrier.

SUMMARY

In this chapter you first learned about frequency modulation. With this type of modulation, the modulating signal causes the carrier frequency to change equally above and below the center frequency. The amount the frequency changes is referred to as frequency deviation. The frequency of the modulating signal determines how fast the frequency deviates, while the amplitude of the modulating signal determines how much the carrier deviates.

Phase modulation, or indirect FM, causes the carrier's instantaneous phase angle to change in proportion to the modulating signal's amplitude. Phase modulation differs from frequency modulation in that with phase modulation, the carrier frequency deviates in proportion to the modulating signal's phase change and its amplitude.

Instead of measuring the amount of modulation in percent, as with AM, angle modulation is measured by the modulation index. It is determined by dividing the frequency deviation by the modulating frequency. Angle modulation can also be expressed as its deviation ratio, which is the ratio of the maximum deviation to the maximum audio frequency.

Angle modulation has an infinite number of sidebands as apposed to AM, which has only two. The waveform's amplitude remains constant as the number of sidebands, their amplitude, or distribution change. The modulation index determines the number of sidebands and their amplitude.

The phase of an FM signal remains constant with modulation, while the carrier frequency changes. With phase modulation, the phase changes abruptly with modulation, while the frequency remains fairly constant.

SELF-TEST REVIEW

1. Angle modulation includes both _____ and _____ modulation.

2. List the three characteristics of frequency modulation.

   1. _____
      _____
   2. _____
      _____
   3. _____
      _____

3. In Phase modulation, maximum deviation occurs when the modulating signal _____ .

4. Phase modulation is also called _____
   _____ .

5. What is the primary difference between PM and FM? _____
   _____
   _____ .

6. The output of the circuit shown in Figure 4-8 is (phase, frequency) modulation. _____ .

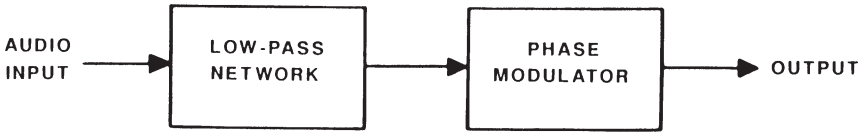


Figure 4-8

Is the output FM or PM?

7. What is the modulation index when a carrier is deviated 35 kHz by a 7 kHz signal? (refer back to Figure 4-5.) _____ .

8. What is the deviation ratio? _____

_____

_____.

9. A 100 MHz carrier is deviated 70 kHz by a 10 kHz audio signal. Using the frequency spectrum in Figure 4-9 and the table in Figure 4-5, determine the modulation index, the frequencies of the sidebands, the amplitude of the sidebands and carrier, and the bandwidth.

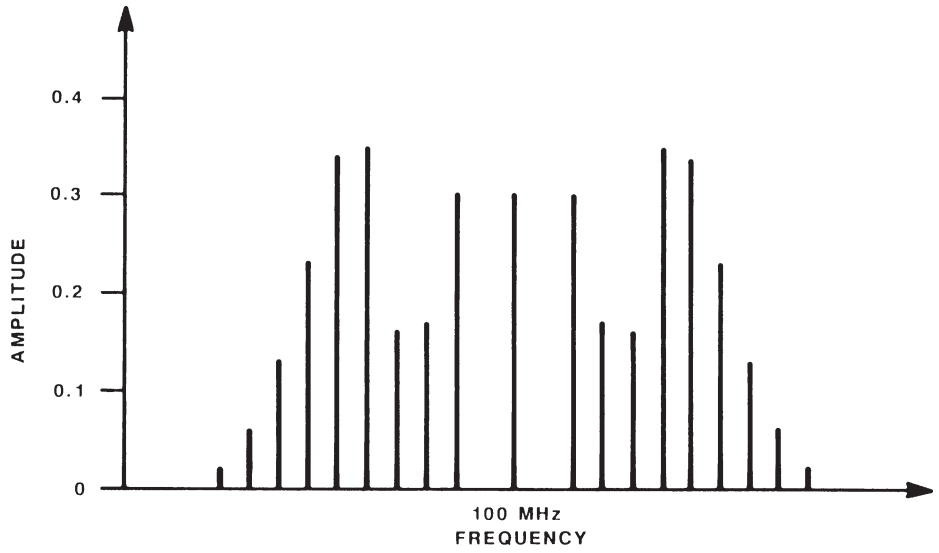


Figure 4-9

Frequency spectrum for problem 9.

10. In what application area is PM most often found? _____

_____

_____.

CHAPTER 5 — The Communication System

INTRODUCTION

In the first four chapters of this course, we discussed the need for modulation and the characteristics of the two basic types of modulation. This chapter is a brief overview of the communications system itself.

The basic block diagram of a communication system is shown in Figure 5-1. It consists of an input device, a transmitter, a transmission channel ,which introduces noise loss and distortion, a receiver, and an output device. The following paragraphs discuss each of these blocks.

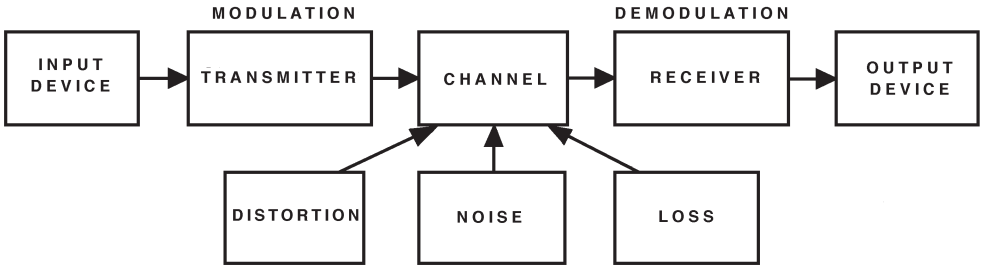


Figure 5-1

A communications system.

OBJECTIVES

When you complete this chapter, you will be able to:

- List the parts of a basic communication system.
- Define transducer and noise.
- Name the four types of noise.
- List the major radio frequency bands.
- State the function of a transmitter, a receiver, and a channel.

INPUT/OUTPUT DEVICES

Input/output devices are transducers, which is the name given to any device that converts an input from one physical form into an output of another physical form. There are many types of input/output devices found in communications systems, and each of them has been developed to meet a particular need for moving information.

Let's look at a few common examples. In broadcasting systems, microphones convert sound waves into electrical signals that are applied to the transmitter. The received signals are converted back to sound waves by the loudspeakers in your receiving set. If you are watching television, the camera in the studio converts light into electrical signals which are transmitted, received at your house, and converted into light by the cathode-ray tube, or picture tube, in your television receiver.

Your telephone handset contains two transducers; there is a microphone which changes the sound waves produced by your voice into electrical signals, and there is also a miniature loudspeaker which converts the electrical signals received from the party on the other end to sound waves that you can hear. In the base of the telephone are other transducers which perform functions that are vital to establishing and maintaining communication. The bell converts an electrical signal into a sound to alert you of a call. The dial or pushbuttons convert physical motions into electrical signals that tell the telephone system how to route your call to the party you want to speak to. And the hook switch, found under the buttons that pop up when you lift the handset, tells the telephone system whether your phone is engaged or is free to accept an incoming call. The "simple" telephone set is actually a sophisticated collection of transducers.

Communication doesn't stop with spoken words. Many input/output devices take written words and convert them into electrical signals for transmission, and then convert them back to written form at the receiver. Examples of such devices are teletypewriters, facsimile machines, and chart recorders.

Probably the most rapidly expanding class of input/output devices found in communications systems is the computer. You see them in stores and at airline reservation desks; you may even have one in your home. Most computer terminals communicate with another computer at least some of the time. In a computer, transducers such as keyboards and storage devices convert movements of your hands or magnetic signals stored on tapes or disks into electrical signals that stand for letters, words, and graphic representations of things which are then

sent by communication systems to another computer. Transducers in the receiving computer convert these signals back into forms meaningful to you, or to the circuits in the receiving computer. And all of it was done with input/output devices of one sort or another.

Many input/output devices are worthy of a complete course of study on their own. You will learn more about some of them later in this course. But they all share one common feature: they produce an electrical signal for use by a transmitter, or they accept an electrical signal from a receiver.

TRANSMITTER

As you have seen, it is usually impractical to transmit the input signal in its original form, or at its original frequency. A transmitter is used to convert the information signal into a form that can be transmitted from the point of origin to where it is needed. It does this by modulating a suitable carrier signal with the input signal, thereby translating the information signal to a higher frequency that is more easily sent from place to place.

Thus, a transmitter must contain a carrier generator and a modulator at a minimum. It may also use amplifiers to increase its power output to help overcome channel losses, noise, and distortion. It also frequently contains circuits that limit the characteristics of the information signal to improve the efficiency and effectiveness of the transmission, such as the band-limiting we discussed earlier to minimize interference.

CHANNEL

The channel is the path over which the information signal is actually sent from the transmitter to the receiver. It may take a number of forms: physical wires, radio links, light-wave transmission, or a combination of these. The requirements of the communication problem determine the characteristics of the channel.

What does that mean? It means that systems are usually designed to use the most cost-effective means of getting information from one place to another. Let's consider the example of a local radio station. It solicits advertisements from local merchants, and airs news and public service announcements of interest to the audience in the local area. It wouldn't make a lot of sense for that station to choose a communication channel that enabled it to be heard nationwide, would it? So,

local radio stations operate in a portion of the radio frequency spectrum where their signals don't usually carry more than a few tens of miles. That restricts their coverage to the audience they want to serve. It also means that the frequency they use can be reused by another local station located outside their coverage area. The same is true of television stations.

On the other hand, national networks want to be able to broadcast the same news program at the same time all over the country, and perhaps all over the world. The radio or televisions that carry the broadcast serve only a local area, but they receive the news program from the network by way of satellites or long-distance telephone systems that make use of radio links and light pipes which can carry the signal to be broadcast thousands of miles without significant degradation. As a result, you can see what happened in China today on the same news program that shows you what happened at the local school board meeting. The items of wide-reaching interest traveled by way of wide-reaching communications systems; the local items, by way of a local communications network.

Communication systems tend to be quite flexible. An excellent example is the telephone system. If you place a local call, say to your neighbor down the street, chances are good that the electrical signals from your telephone and your neighbor's travel on wire cable all the way from your house to the telephone central office and back to your neighbor. However, if you call Aunt Flo, who lives 1,500 miles distant, your call probably traveled on wire cable to the central office, then on several radio, satellite, and/or light wave systems to an office near Aunt Flo, and then over wire cable once again to her telephone. The system chose channels to use based on the characteristics of each of your calls.

The channel is the key to communication systems. Since you know that information is rarely sent at its original frequency, but is instead moved up in the frequency spectrum, you also know that the channel must operate at a frequency higher than the transmitter input signal. Indeed it does. Figure 5-2 shows the radio frequency spectrum, and identifies where some of the common channels of communication systems are to be found. Notice that there is no such thing as a "typical" channel frequency. Channel frequencies are chosen by systems designers to provide the characteristics needed to satisfy the requirements of the communication system being designed. You will see later that greater bandwidth, and therefore a greater rate of information transfer, can be achieved at high carrier frequencies than at low frequencies. This is a major reason why many communication systems operate at high channel frequencies.

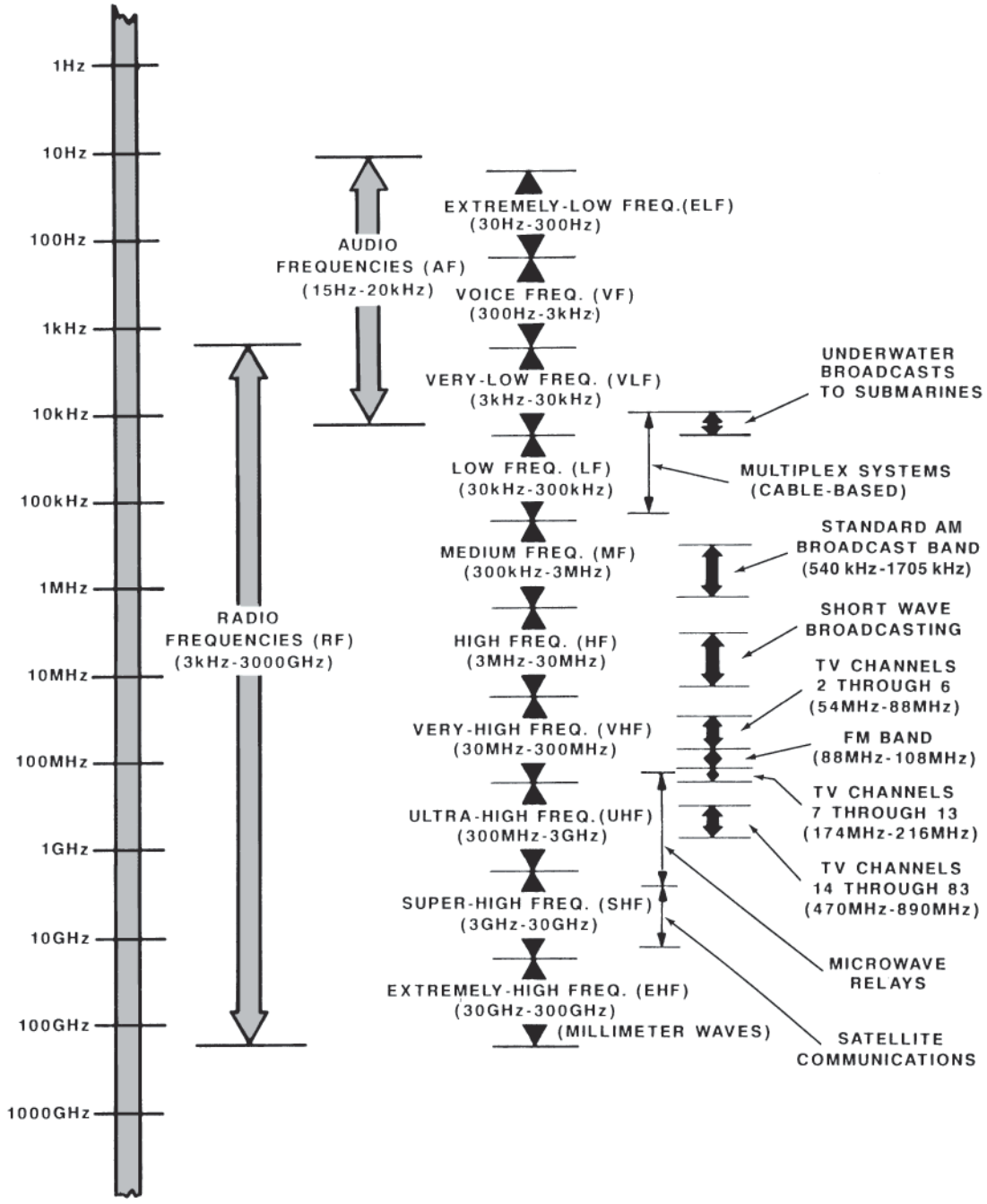


Figure 5-2

The radio frequency spectrum.

NOISE

Unwanted random signals which interfere with the reception of wanted signals are called noise. In Figure 5-1, only one source of noise is shown for simplicity; there are actually four basic types of noise. These are atmospheric noise, space noise, man-made noise, and internal noise.

Atmospheric noise is caused by electrical disturbances in the earth's atmosphere. The most common source of this noise is lightning, which generates radio energy across a broad band of frequencies every time it discharges. Radio waves can propagate around the world given the right conditions, and somewhere in the world there is always a lightning storm in progress. Therefore, this source of noise is always with us, even when there is no local lightning storm.

Space noise is caused primarily by radiation from the sun and the other billions of stars in the universe. We do not yet understand all of the sources of radio signals in the universe, but the most common and most powerful sources are the stars. Our sun is the nearest star to us, and constantly throws off energy across a wide frequency spectrum. In addition to the steady state noise produced by the sun, there are also transient storms and disturbances (called sunspots and solar flares) which produce even higher levels of noise from time to time. All the stars in the universe generate noise in this way; the noise from the sun is most troublesome because the sun being relatively close to the earth is also the loudest.

Almost all of us are familiar with man-made noise. The electric motor of an appliance can generate wideband radio noise that can virtually wipe out radio or television reception. So can automobile ignitions, high-voltage power line leakage, fluorescent lights, industrial equipment, and other sources too numerous to count.

Internal noise is generated by the devices inside electronic equipment itself. Internal noise can be divided into two types: thermal noise and shot noise. Thermal noise is generated in any resistance above absolute zero in temperature, and is caused by the random motion of electrons in the resistance. This random motion generates small currents which flow randomly, causing a signal that can interfere with the one we want. Shot noise is caused by the random variations that occur in the instantaneous rate of arrival of electrons at the collector of a transistor or the drain of an FET, which also cause minute random variations in the average current of the desired signal. Noise is a problem in all communication systems. If not properly dealt with, it can render a system useless.

DISTORTION

Perfect communication channels do not exist. Each channel has characteristics which distort the signal that travels through the channel. You have already seen that transmitters and receivers cannot accommodate infinite bandwidth. This is one source of distortion. Electronic devices, such as transistors and integrated circuits, do not perfectly reproduce at their outputs the signals that were applied as inputs. This, too, is a source of distortion.

Proximity to other communication channels, physically or in frequency, provides other sources of distortion. You have probably heard strange noises, squeals, voices, and "Donald Duck" sounds at times during telephone calls or radio and television broadcasts. All of these are caused by one form or another of distortion in the communications channel. Later, we will look at distortion in more detail to see how best to deal with it.

LOSS

The energy in the signal generated by the transmitter is rapidly dissipated as it travels through the communication channel. There is no loss-free transmission medium. Wires have resistance which convert electrical energy to heat. Radio signals are broadcast into space, and only a small portion of the radiated energy is intercepted by antennae for use by receivers. Fiber-optic cables ("light pipes") introduce considerable loss to the passage of light waves.

Loss performance is important in the analysis and selection of the channel, because the signal at the receiver must be strong enough to be properly reconstructed as an accurate image of the transmitted signal. Some channels have more loss than others, but all channels have some loss, and all signals are much weaker at the receiver than they were when they left the transmitter. As an example, large local radio stations typically broadcast at a power level of 50,000 watts; the signal at your receiver, however, is measured in millionths of a watt. The loss in other channels is many times more than this.

RECEIVER

By the time the transmitted signal reaches the receiver, most of its power has been dissipated, and distortion and noise have been added to it. The receiver must amplify this weak, distorted signal enough to overcome the internal noise generated in the receiver, and must separate it from the other transmitted signals present in the same frequency band. It must then demodulate the signal to recover the original information. The receiver may have to process the demodulated signal to remove some of the effects of distortion and noise. This is a very difficult task. Receiver design has been developed over many years to provide extraordinary performance in the reception of very weak signals. In later chapters of this course, we will look in detail at receivers designed for particular types of channels.

SUMMARY

In this chapter, you reviewed the components and characteristics of communications systems, paying attention to some of the more common things that upset the perfection of communication: noise, distortion, and loss. In the following chapters, you will build on this review to learn more about amplitude modulation.

SELF-TEST REVIEW

1. List the parts of a basic communications system. _____

   _____

   _____

2. What is a transducer? _____

   _____

   _____

3. What are the names and frequencies of the major radio frequency bands? _____

   _____

4. What is noise? _____

   _____

5. List the four types of noise. _____

_____

6. What is the function of a transmitter? A receiver? _____

_____

_____

Section 2

Amplitude Modulation

CHAPTER 6 — AM Basics

INTRODUCTION

To obtain amplitude modulation, the information signal and the carrier must be combined in a stage called the modulator. Low-level AM modulators are relatively simple. However, AM transmitters must produce power levels ranging from 5 watts, for a citizen's band (CB) transceiver, up to 2 million watts, for a broadcast transmitter. Such high power levels require very complex modulators. This chapter discusses both types of modulators and the characteristics of an AM signal.

OBJECTIVES

When you complete this chapter, you will be able to:

- Name five methods of generating amplitude modulation.
- Define "modulator," and describe how modulation occurs.
- Find the required modulator power when given input power and modulation percentage.
- State the advantages and disadvantages of collector modulation, series modulation, and progressive series modulation.

CREATING AMPLITUDE MODULATION

The following sections describe five ways to generate amplitude modulation.

The Diode Modulator

A simple circuit for producing amplitude modulation is shown in Figure 6-1. The modulating signal, in this case audio, is applied at the top of R1 while the carrier is applied at the top of R2. The signal at the junction of R1 and R2 is the sum of the carrier and audio. Notice that the carrier is not amplitude modulated at this point. It is simply added to the audio signal.

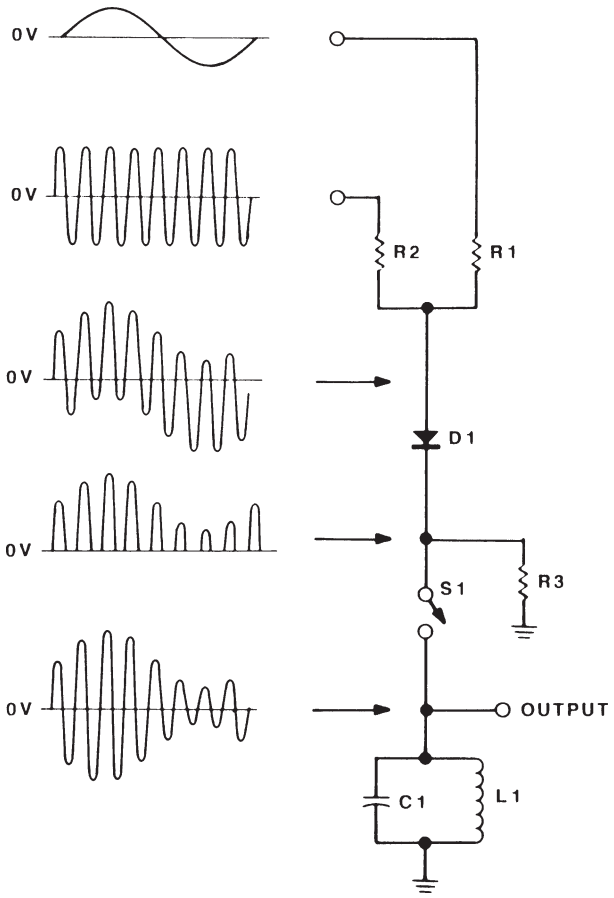


Figure 6-1
The diode modulator and its waveforms.

When S1 is open, D1 conducts through R3 when the signal on the anode swings positive. However, when the signal swings negative, D1 cuts off. Thus, the signal developed across R3 consists of positive-going pulses as shown. Notice that the positive pulses vary in amplitude in accordance with the amplitude of the audio signal.

S1 is included in the circuit merely for explanation purposes. Normally, it is omitted and the tank circuit composed of C1 and L1 is connected directly in parallel with R3.

The purpose of the tank circuit, which is tuned to the carrier frequency, becomes clear when S1 is closed. Each time D1 conducts, a pulse of current flows through the tank. This causes the tank to resonate and the flywheel action of the tank produces a negative half-cycle for each positive input pulse. The high amplitude positive pulses cause high amplitude negative pulses, and the low amplitude positive pulses cause low amplitude negative pulses. Therefore, each negative half-cycle will have the same amplitude as the positive half-cycle. The output is an AM wave. Thus, this simple circuit produces amplitude modulation.

The diode modulator is an example of nonlinear mixing. That is, the audio and carrier signals are mixed across the nonlinear resistance of D1. The result of this mixing action is sum and difference products. Therefore, the signals across D1 are the original frequencies and their sum and difference. If the carrier is 100 kHz and the audio signal is 10 kHz, the sum is 100 kHz + 10 kHz = 110 kHz, and the difference is 100 kHz – 10 kHz = 90 kHz. These are the sideband frequencies.

> *Nonlinear mixing* occurs when the audio and carrier signals are mixed across the nonlinear resistance of a diode.

The 100 kHz carrier and 10 kHz audio signal are also present. Since the resonant circuit of C1-L1 is tuned to 100 kHz, only the carrier and sidebands pass to the output. The 10 kHz audio signal is filtered out. Thus, the output of the diode modulator is a true amplitude modulated signal.

The Absorption Modulator

The absorption modulator is used primarily in VHF, UHF, and higher frequency signal generators. Most absorption modulators use PIN diodes because of their variable impedance characteristics. Usually, several PIN diodes are arranged in an attenuator network. This arrangement not only allows amplitude modulation, but also variable output level control, which makes it an extremely versatile device for use in signal generators.

Another advantage of the absorption modulator is that it does not affect the voltage applied to the frequency generating device or circuit. This is very important at VHF, UHF, and higher frequencies since any voltage changes at the oscillator may cause undesired frequency variations.

You will recall that a PIN diode is formed by placing an intrinsic (I) semiconductor material between the P-type and N-type materials of a junction diode. Above 100 MHz, this causes the diode to act like a linear resistance by conducting in both directions. However, its resistance depends on the amount of forward bias current. Therefore, the PIN diode acts as a current-controlled resistance. This makes it ideal for attenuator networks.

Figure 6-2 shows a typical pi-section attenuator. The output signal level is determined by the values of the three resistors. When PIN diodes are used as the variable resistances, the output level is set by varying the DC bias current applied to the PIN diodes, thereby establishing an output signal level. To modulate the output signal, the modulating signal is added to the DC bias, just like the two signals were added in Figure 6-1, and this varying bias voltage is applied to the resistances. The result is an amplitude modulated RF carrier.

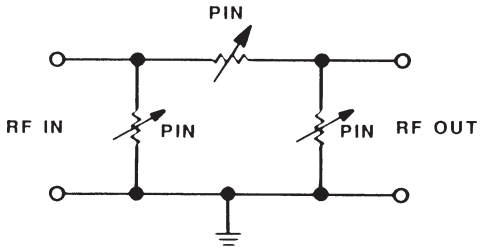


Figure 6-2

A simplified diagram of a PIN diode absorption modulator.

If 100% modulation is to be possible, the attenuator must be set to provide an average output signal of approximately one-half the available output power from the carrier source. That way, equal variations in carrier strength above and below the average output level are possible, as is required for 100% modulation.

The efficiency of the absorption modulator is very low. However, at microwave frequencies it is one of the few methods of amplitude modulation available. Moreover, since it is normally used at low power levels, efficiency is not an important requirement.

The Collector Modulator

The most commonly used AM technique is collector modulation or, in vacuum tube modulators, plate modulation. In collector modulation, the modulating signal is applied to the transistor's collector in series with the DC supply voltage.

A typical collector modulation circuit is shown in Figure 6-3. Q1, and its associated circuitry, form a class C RF power amplifier.

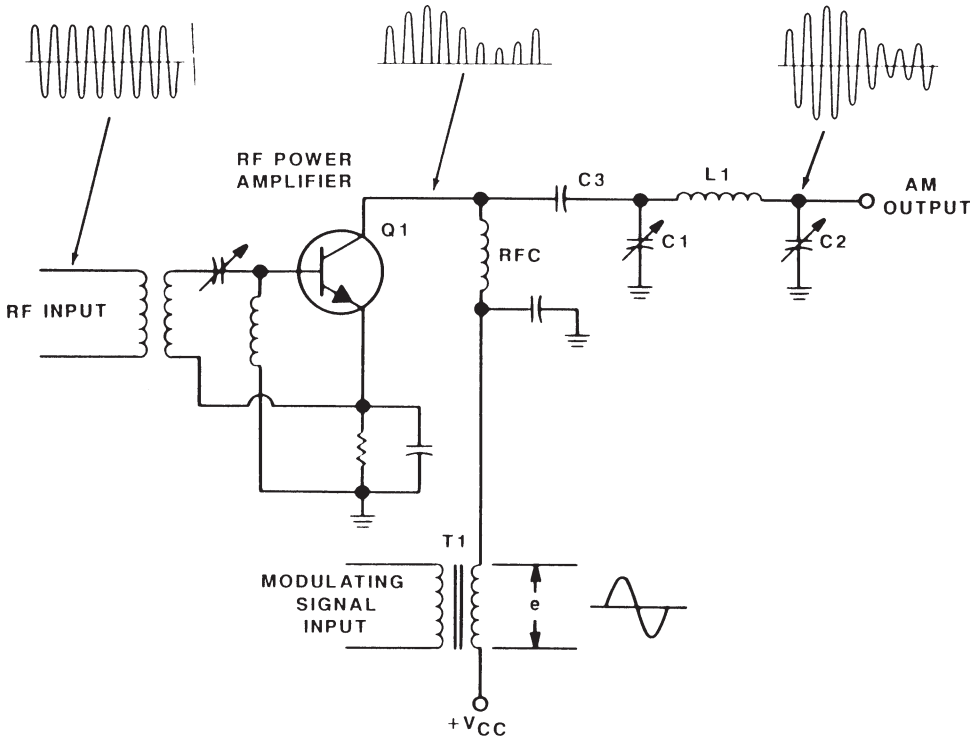


Figure 6-3

Basic collector modulation circuit.

The modulating signal is applied to the primary of transformer T1. As you know, a voltage is developed across the secondary winding of a transformer when a signal is applied to the primary. In this circuit, the secondary of T1 is connected in series with the DC collector supply voltage, V_{cc}. So, the voltage applied to the collector of the transistor Q1 at any moment is the sum of the voltage across the secondary of T1, e, and V_{cc}.

When no modulating signal is present, there is only a small, constant voltage drop across the secondary of T1 due to the resistance of the secondary winding. Therefore, virtually the full value of V_{cc} is applied to the collector through the secondary of T1 and the radio-frequency choke, RFC. However, when the T1 secondary voltage, e, swings positive, it adds to V_{cc}, increasing the voltage applied to the collector, which increases the collector current flow, and thereby increases the amplitude of the output signal. Conversely, when the modulating signal swings negative, voltage e subtracts from V_{cc}, lowering the instantaneous collector voltage and thus the amplitude of the output signal.

In this way, the peaks and troughs of the modulating signal cause corresponding peaks and troughs in the amplitude of the RF output signal, creating an amplitude modulated signal.

Since Q1 is a class C amplifier, only positive pulses appear at its output. To form a complete AM waveform, these output pulses must be applied to a resonant or "tank" circuit. The circuit of L1, C1, and C2 in Figure 6-3 form the output tank for this circuit. You may recognize this circuit as a pi network. A similar circuit is often used as a low-pass filter in power supplies. The circuit shown here is also a low-pass filter, but it is designed to be different from the power supply filter in two important ways: it is a resonant tank circuit, and an impedance matching network. Let's examine its operation in detail.

Figure 6-4 shows the basic output network. The output signal of Q1 is developed across the RF choke (RFC) and capacitively coupled to the pi network by C3. If the schematic is rearranged as shown in Figure 6-4B, you can see that the combination of C1 and C2 are in parallel with L1, creating a parallel-resonant circuit. Therefore, by tuning C1 and C2, the circuit will resonate at the output frequency. Figure 6-4B also shows that the input is across C1, while the output is taken across C2. By varying the ratio of C1 to C2, and impedance match can be obtained between Q1 and the output load. Thus, not only is the pi network a low-pass filter, but it is also a resonant tank circuit and impedance matching network.

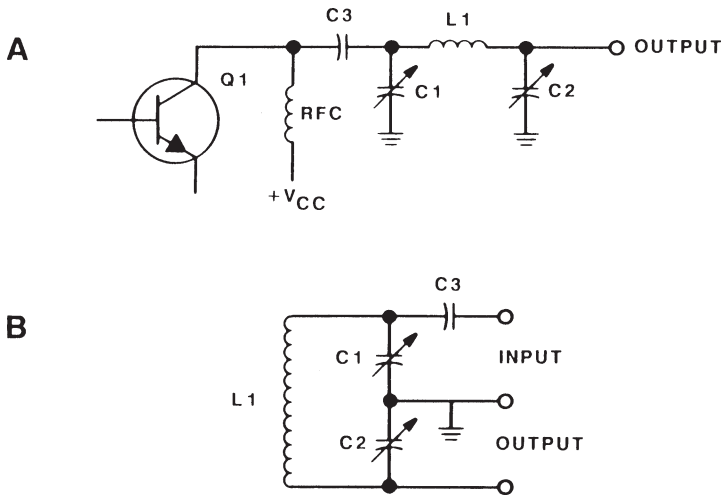


Figure 6-4

Pi output tank circuit.

One problem with the collector modulation circuit of Figure 6-3 is that 100% modulation is not possible if the transistor saturates before reaching the positive peak of modulation. Also, at the negative trough, the base RF drive feeds through the base-collector junction capacitance, thereby preventing zero output at the negative peak. Both of these problems can be overcome by modulating both the driver and final amplifier. This is shown in Figure 6-5. By modulating the driver, extra drive is provided to the final amplifier at the positive peak, allowing it to reach a full 100% modulation. At the negative trough, the drive is reduced sufficiently to prevent RF feedthrough.

The modulator in the circuit of Figure 6-5 is a push-pull audio amplifier. It must deliver at least one-half of the input carrier power to obtain 100% amplitude modulation. Therefore, a relatively high power audio amplifier is required. The push-pull circuit delivers high power at much better efficiency than a single-ended class A amplifier. Let's examine exactly what the modulator power requirements are.

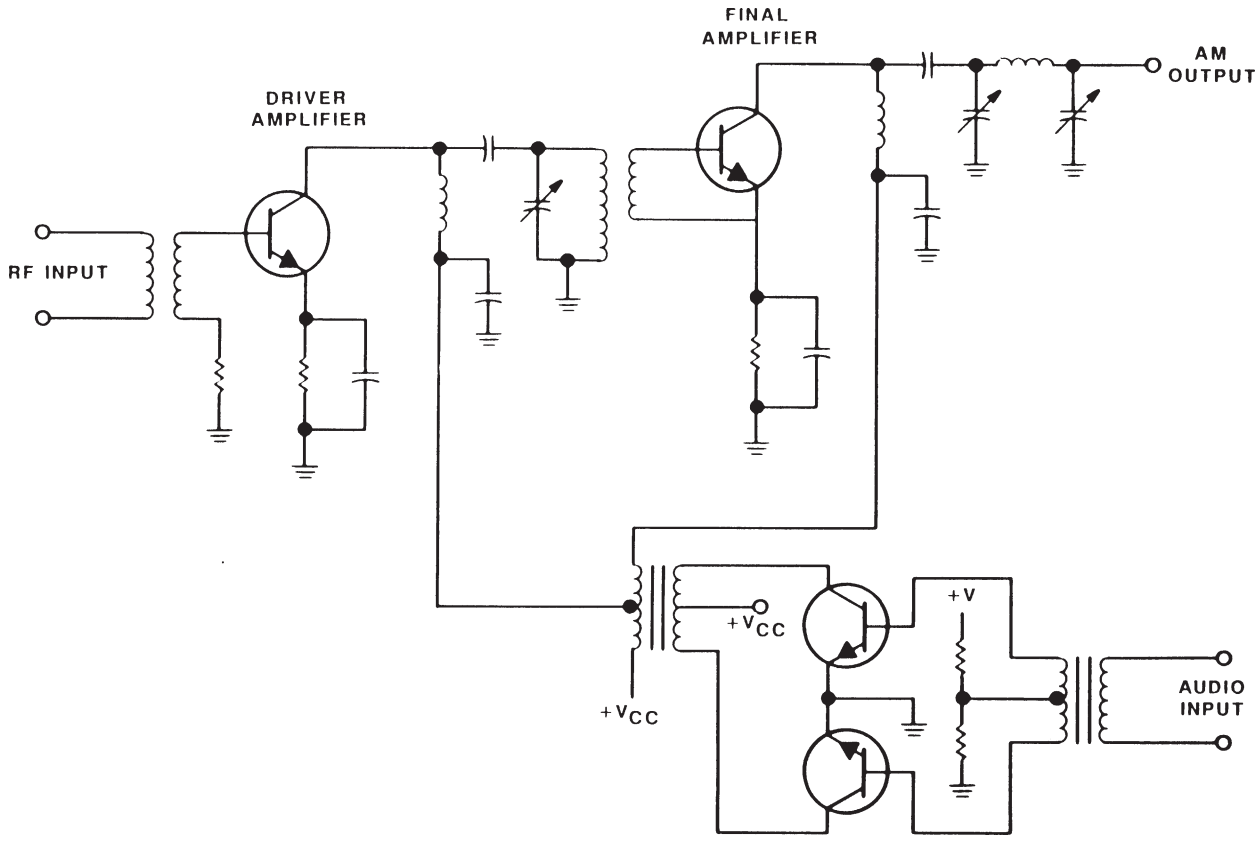


Figure 6-5

Collector modulation of both driver and final amplifiers.

As an example, suppose that V_{cc} is 28 volts and I_c is 1 ampere. The DC input power is:

$$P_{in} = V_{cc} \times I_c$$
$$P_{in} = 28\ V \times 1A$$
$$P_{in} = 28\ W$$

The impedance of the collector circuit is:

$$Z_c = \frac{V_{cc}}{I_c}$$

$$Z_c = \frac{28\ V}{1\ A}$$

$$Z_c = 28\Omega$$

This is the impedance that the audio modulating power must be developed across.

To modulate the amplifier stage 100%, the sum of the instantaneous audio peak voltage and V_{cc} must be twice the V_{cc} for the positive peak and zero for the negative trough. Since V_{cc} is 28 V, the modulator must double this by supplying 28 V peak to the transformer secondary. This voltage must be developed across the 28Ω collector impedance. To find the power required, you can use the formula for power. However, this equation applies only to effective or rms values of AC voltages. Therefore, the peak audio voltage must be converted to rms:

$$E_{rms} = E_{pk} \times 0.707$$

$$E_{rms} = 28\ V \times 0.707$$

$$E_{rms} = 19.8\ V$$

Now, using the power formula:

$$P = \frac{E^2}{R}$$

$$P = \frac{19.8\ V^2}{28}$$

$$P = 14W$$

Thus, the modulator output must be 14 W to obtain 100% modulation of a 28 W RF amplifier.

From these figures you can also determine the sideband and carrier power in the transmitted signal. Continuing our example, if the efficiency of the final amplifier is 75%, the carrier output power is:

$$P_{out} = P_{in} \times E_{pp}\%$$

$$P_{out} = 28\ W \times 75\%$$

$$P_{out} = 28\ W \times 0.75$$

$$P_{out} = 21\ W$$

At 100% modulation, the modulator supplies 14 W to the final amplifier. The amount of modulator power that appears at the output is also determined by the final amplifier efficiency. Therefore,

$$P_{out} = P_{in} \times E_{pp}\%$$

$$P_{out} = 14\ W \times 0.75$$

$$P_{out} = 10.5\ W$$

The modulator power supplied to the output, which is divided equally between the sidebands, is 10.5 W or one-half the carrier output power.

An equation for determining modulator power or sideband power for any modulation percentage is:

$$P_m = \frac{m^2 P_c}{2}$$

Where: P_m = modulator power or sideband power

m = percent of modulation

P_c = carrier power.

To determine modulator power, P_c must be the carrier **input** power. To find sideband power, P_c must be the carrier **output** power.

Let's try a couple of examples. What is the sideband power for a 100-watt carrier at 50% modulation?

$$P_m = \frac{m^2 P_c}{2}$$

$$P_m = \frac{0.5^2 \times 100W}{2}$$

$$P_m = 12.5 \text{ W}$$

Since sideband power is 12.5 W, the power of each sideband is 6.25 W.

What modulator power is required for 80% modulation of an RF amplifier with an input power of 1000 W?

$$P_m = \frac{m^2 P_c}{2}$$

$$P_m = \frac{0.8^2 \times 1000W}{2}$$

$$P_m = 320 \text{ W}$$

Series Modulator

While collector modulation is the most commonly used technique, it does have disadvantages. The most serious of these is the modulation transformer itself. It is large, heavy, costly, and introduces frequency distortion. That is, due to the inductance of the transformer, some frequencies are attenuated more than others. One method that eliminates the modulation transformer is series modulation. This technique uses a transistor in place of the transformer.

A basic series modulator is shown in Figure 6-6. In this circuit, the current through the final amplifier, Q1, is controlled by the series modulator, Q2. Therefore, any increase in current flow through Q2 increases the output of Q1. Likewise, any decrease in Q2 current flow decreases the output of Q1. Thus, the series modulator amplitude modulates the RF signal present at Q1 and it does it through direct coupling rather than transformer coupling. This minimizes frequency distortion, and replaces the expensive transformer by a relatively inexpensive transistor.

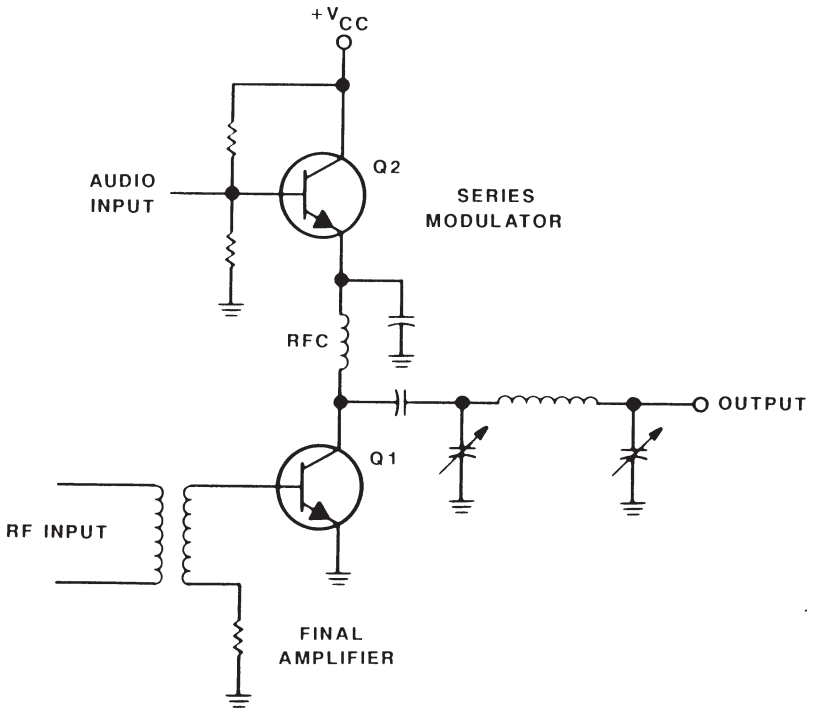


Figure 6-6
A basic series modulation circuit.

Series modulation has one significant disadvantage: it is very inefficient. You may recall that at the positive peak of 100% modulation, the voltage at the final amplifier must be doubled. As an example, if the final amplifier has a collector voltage of 100 V and collector current of 1 A, the input power is 100 W. In order to 100% modulate this amplifier, the collector voltage and current must double for the positive peak of modulation. This additional voltage must be dropped across the series modulator at 0% modulation. Therefore, at 0% modulation, the series modulator drops 100 V at 1 A. This is shown in Figure 6-7. Note that the modulator is dissipating a full 100 W. This is very inefficient operation. Collector modulation is much more efficient, since at 0% modulation, there is almost no voltage dropped across the secondary of the modulation transformer.

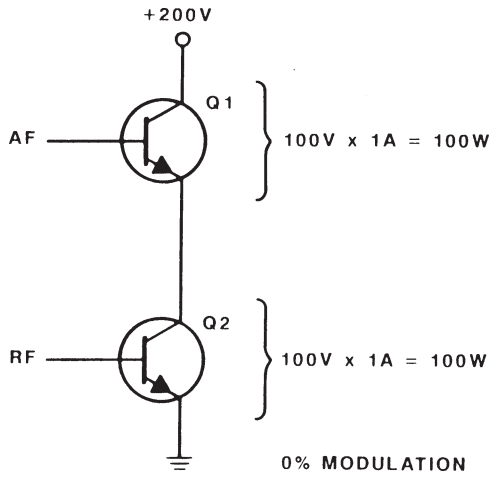


Figure 6-7

Power distribution at 0% modulation in a series modulator.

Progressive Series Modulator

One circuit that greatly improves the efficiency of series modulation while still maintaining the advantage of direct coupling is the progressive series modulator. This circuit is shown in Figure 6-8. It uses two modulator transistors and two power supplies in a special switching configuration. At 0% modulation and during the negative troughs of modulation, Q1 is turned off and, therefore, the 104 V supply is disconnected from the circuit. During the negative troughs, Q2 controls the current and voltage applied to the final amplifier. For positive modulation, Q2 saturates and Q1 turns on. As soon as Q1 turns on, D1 is reverse biased and, therefore, disconnects the 52 V supply. At this time, Q1 controls the voltage and current to the final amplifier. The 104 V supply is now connected, which allows a full positive peak of modulation.

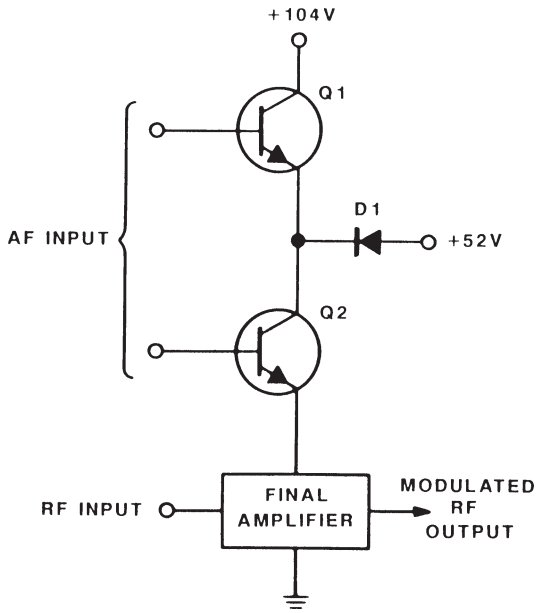


Figure 6-8

A simplified diagram of a progressive series modulator.

The advantage of this arrangement is that, at carrier-only conditions or 0% modulation, very little power is dissipated across the modulator. This is because, at 0% modulation, Q1 is off while Q2 is on and saturated. Its saturation voltage is approximately 2 V, therefore, this leaves 50 V from the 52 V supply for the final amplifier. Thus, at carrier-only conditions, 50 V is applied to the final amplifier. If the current is 2 A, input power is 50 V × 2 A = 100 W. However, the power dropped across modulator transistor Q2 is only 2 V × 2 A = 4 W. This is a significant difference over the normal series modulator. Thus, the progressive series modulator offers much better efficiency while maintaining direct coupling. The only disadvantage is increased circuit complexity.

AM CHARACTERISTICS

Now that you have learned how to create an AM signal, it is time to examine the characteristics of the signal.

Power and Efficiency

You learned earlier that most of the transmitted signal power is found in the carrier. By examining the equations used to determine how much power is found in the sidebands, you can see that half the transmitted signal power is found in the carrier at 100% modulation level. When the modulation percentage is less than 100%, even more of the power is in the carrier. The remaining power is split evenly between the upper and lower sidebands. So, the maximum amount of signal power found in either of the sidebands is 25% of the total transmitted power.

So what? Remember that the carrier doesn't vary along with the information signal; the carrier is constant in both frequency and amplitude. Therefore, it carries no information. The carrier is necessary only to create the sidebands, where the information is carried, and to detect the sidebands at the receiver. Along the transmission path, the carrier serves no useful purpose, yet it consumes the lion's share of the transmitted power. This is not an efficient way of sending information, but it is an essential characteristic of an AM signal.

Bandwidth

In Chapter 3 we briefly discussed the bandwidth of an AM signal. It is important to remember that the bandwidth is always twice the highest modulating frequency. If the highest frequency applied to an AM modulator is 5 kHz, then the bandwidth of the transmitted signal is 10 kHz. This 2-to-1 ratio between the modulating signal frequency and the signal bandwidth is another basic characteristic of amplitude modulation.

Phase Relationships

In generating an AM signal, the phase relationships between the carrier and the modulating signal are important. These relationships are preserved in the modulated signal. So, for an AM signal to be received and demodulated properly, it is necessary that the carrier and each of the sidebands be received exactly as they were transmitted without variation in phase or amplitude.

Unfortunately, most transmission paths do not provide for such perfect signal transmission. Since the carrier, the upper sideband, and the lower sideband are each at a separate frequency, they are frequently affected differently as they pass through the transmission channel. At the receiving end, this difference equates to distortion in the received signal.

Noise and Distortion

The information in an AM signal is conveyed by variations in the amplitude of the signal. Unfortunately, most sources of unwanted noise, like lightning, also generate signals which have varying amplitudes. You could say that noise tends to be amplitude modulated, rather than angle modulated.

You know that voltages occurring across the same resistance add together. The amplitude of a signal is represented by its voltage, and the transmission path contains resistance. When noise occurs along the path of an AM transmission, the noise signal is added to the AM signal. Once the noise is added to the desired signal, it's impossible to remove it completely, because the noise is random. The result is a signal which contains unwanted amplitude variations which make it much more difficult to recover the desired signal. Vulnerability to noise is also a basic characteristic of AM.

Not all of the characteristics of AM are desirable in the design of a communications system. You would prefer to have an efficient signal, which occupied minimum possible bandwidth, and was not particularly susceptible to disruptions such as distortion and noise. In the next chapter, we will examine some ways to improve the AM signal for practical communications.

AMPLITUDE MODULATION, PULSES, AND NOISE

So far, we have always considered that the signal modulating the AM carrier was some form of continuous wave: a single sine wave, or voice signals. This need not be so. A carrier can be modulated by pulses just as easily as by a continuous waveform. However, when pulse modulation is used, it is wise to think about the form of the resulting modulated wave.

Consider modulating a 100 kHz carrier with a 1 kHz square wave. You know that the square wave can be represented by an infinite series of sine waves, each one an odd harmonic of 1 kHz. For the purposes of this illustration, we will pass the 1 kHz square wave through a low-pass filter, as shown in Figure 6-9, that removes all harmonics above 6 kHz.

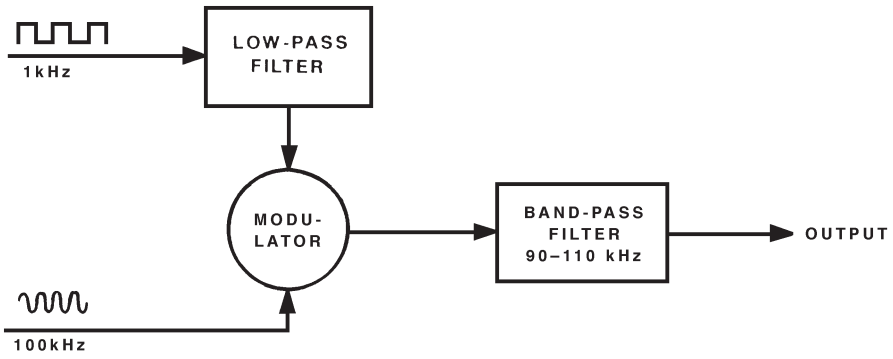


Figure 6-9

100 kHz carrier modulated by band-limited I kHz square wave.

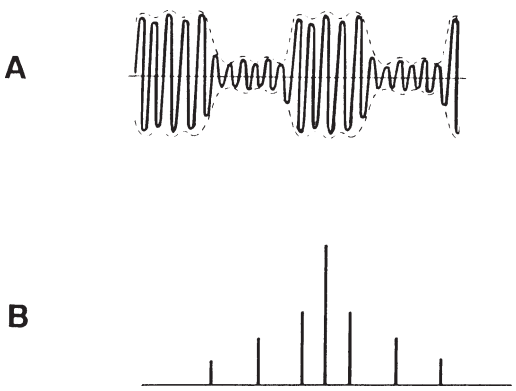


Figure 6-10

Part A shows the time domain view of the signal generated in Figure 6-9. Part B shows the frequency domain view of signal generated in Figure 6-9.

The square wave has no even harmonics in its sine wave representation, so the modulator sees the following sine-wave signals at its audio input: 1 kHz, 3 kHz, and 5 kHz. The modulator performs its mixing function, producing the following output frequencies (the 1 kHz, 3kHz, and 5 kHz signals are removed by the bandpass filter):

105 kHz
103 kHz } Upper sideband
101 kHz

100 kHz — Carrier

99 kHz
97 kHz } Lower sideband
95 kHz

The result of this modulation in both time and frequency domains is shown in Figure 6-10.

Examine Figure 6-10 closely. The time domain view shows what you would expect: a square wave variation in the carrier envelope, corresponding will the envelope of the modulating square wave signal, as represented by its first three harmonics.

But the frequency domain reveals something not so intuitive: this signal, modulated with a 1 kHz square wave, has a bandwidth of 10 kHz, not 2 kHz as you might expect. Also, the sidebands of the modulated signal contain only discrete harmonics of the square wave repetition frequency. There are none of the subtle, varying frequencies found when human speech is used as the modulating signal. In fact, were you to examine the frequency domain representation of this signal without knowing how it had been generated, you might suspect it to be FM.

Only because we limited the bandwidth of the applied square wave is the bandwidth of the modulated signal 10 kHz. Had the square wave been applied directly to the carrier, significant harmonics beyond the fifth would be represented in the sidebands, with an accompanying increase in bandwidth.

What can you learn from this? First, the bandwidth of an AM carrier modulated by pulses is considerably greater than twice the pulse repetition frequency, and is determined in practice by the amount of filtering applied to the pulses before they are input to the modulator. Second, the sidebands of the modulated signal contain discrete sine waves which are the Fourier set of sine waves that represent the particular modulating wave. Because of this, a frequency domain scan of such a signal may not make it obvious if it is AM or not.

You wouldn't normally modulate a carrier with a square wave except for testing purposes, though. As with the sine wave, it doesn't convey any information. But you would very likely use a stream of pulses representing data as the modulating signal. The only practical difference between a pulse stream and the square wave case we just examined is that there are more sine wave components of the pulse stream, so the frequency domain appears somewhat more "crowded."

In order to reconstruct the pulse stream, and thus the data it represents, at the receiver, the AM signal must be received accurately. Even if we are able to successfully counter the problems caused by distortion, however, there remains a major difficulty: noise.

Unlike speech, data is not very redundant. You need to receive all of it correctly to make sense of any of it. Figure 6-11 illustrates what happens to a pulse-modulated AM signal in the presence of noise. Figure 6-11A shows the signal as it is transmitted. Figure 6-11B shows a random noise signal that is added to the AM signal during transmission. Figure 6-11C illustrates the sum of these two signals, and Figure 6-11D shows what the receiver is likely to recover and pass on for processing. Notice the difference between the envelope of Figure 6-11A, which represents the modulating pulse stream, and the detected pulses at Figure 6-11D.

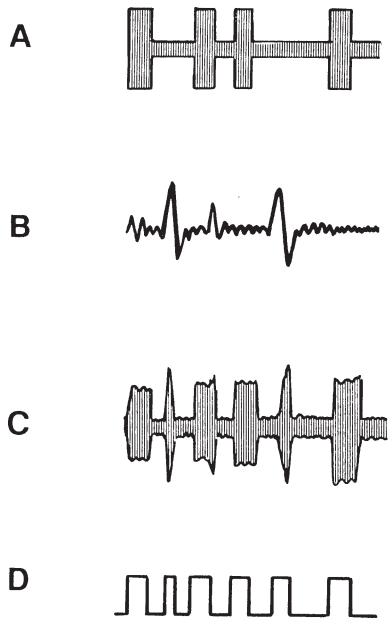


Figure 6-11

The effects of noise on a pulse-modulated AM signal.

There is nothing different in the way that noise affects the AM signal modulated by pulses than the way noise affects all AM signals. The noise, after all, doesn't know what created the signals it encounters. But the effect on the communication of information as a result of the noise, as compared to the effect of a few noise spikes in the midst of a voice conversation, is dramatic. AM, pulses, and noise are a bad mix. It is very difficult to carry on accurate communications using pulse modulation of an AM carrier in the presence of noise.

SUMMARY

In this chapter you learned about the diode, absorption, collector, series, and progressive series types of amplitude modulators. Collector modulators are the most-common type, but require a large, expensive modulation transformer. Series modulators do not require a transformer, but are much less efficient. A progressive series modulator does not require a transformer and is more efficient than the series type, but it's circuitry is more complex.

You learned how to calculate the amount of modulator power that is required when you know the input power and the percentage of modulation. You can also calculate sideband power when you know the carrier power and the percentage of modulation.

You also learned some characteristics of an AM signal: power and efficiency, bandwidth, and phase relationships, and how the signal is susceptible to noise and distortion.

Finally, we considered the special case where data is used as the modulating signal. In this circumstance, we found that using AM over a noisy path is not likely to produce reliable communications.

SELF-TEST REVIEW

1. Name five methods of amplitude modulation.

   _____

   _____

2. What is a modulator?

   _____

   _____

3. What are the three functions of a pi output network?

   _____

   _____

4. How much modulator power is required for 75% modulation of a 10 kW amplifier? _____

5. What is the power in the upper sideband for a carrier of 100 W when modulation is 55%? _____

6. What is the primary disadvantage of collector modulation?

_____

_____

7. List the advantage and disadvantage of series modulation.

_____

_____

8. How does progressive series modulation overcome the disadvantage of series modulation? _____

_____

9. Which of the modulators discussed in this unit use nonlinear resistance to obtain AM? _____

_____

10. Name the four characteristics of an AM signal. _____

_____

_____

CHAPTER 7 — Suppressed Carrier AM

INTRODUCTION

The type of amplitude modulation discussed in Chapter 6 is referred to in this chapter as standard AM. In this system, the carrier and both sidebands are transmitted just as they appear at the output of the modulator. This system is used by standard AM broadcast stations and most citizen's band radios. Its primary advantage is that is uses straightforward and inexpensive transmitting and receiving equipment. However, as you have just seen, it has several disadvantages. One of these disadvantages can be overcome by transmitting a suppressed carrier signal.

OBJECTIVES

When you complete this chapter, you will be able to:

- Explain the characteristics of standard AM.
- Name three types of balanced modulators.
- Identify a double sideband suppressed carrier waveform.
- Explain the difference between low-level and high-level modulation.
- State the primary advantage and disadvantage of suppressed-carrier AM.

BALANCED MODULATORS

One disadvantage of standard AM, wasted carrier power, can be eliminated by suppressing, or removing, the carrier before transmission. This is accomplished by using a balanced modulator. This device mixes the audio and RF carrier. Then, due to its circuit configuration, it allows only the mixing products or sidebands to pass to the output. The resultant signal is called double sideband suppressed carrier. It can be transmitted and received without a carrier and, thus, offers more efficient transmission than standard AM. There are several different types of balanced modulators. Each type has its own advantages and disadvantages. Let's discuss several of the more common balanced modulators.

Twin-FET Balanced Modulator

A relatively simple balanced modulator is shown in Figure 7-1. Here, the RF carrier input is applied in phase to the gates of Q1 and Q2. This causes equal but opposite drain currents to flow in the primary of T2. The result is zero output when the carrier alone is applied to the circuit. Potentiometer R1 is used to compensate for differences in the two FETs. The output is monitored and R1 is adjusted for minimum or zero output with the carrier only applied to the circuit.

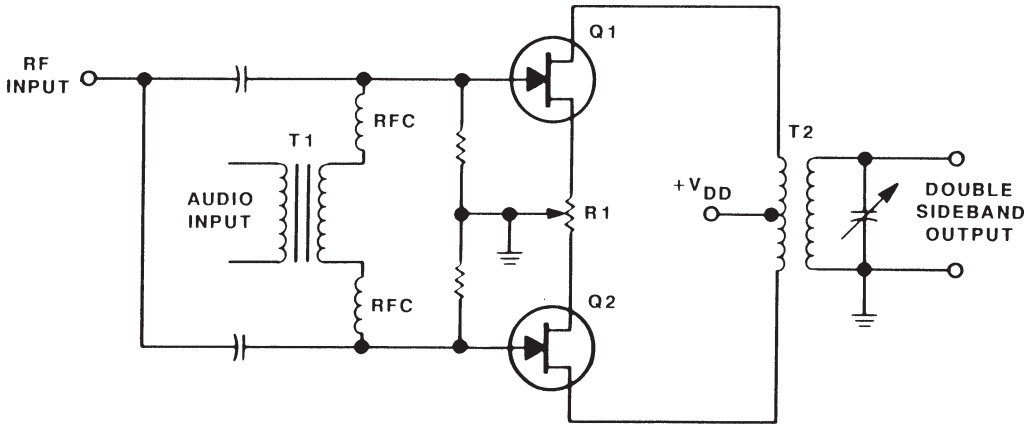


Figure 7-1
The twin-FET balanced modulator.

When audio is present, it is applied to the gate of Q1 180° out of phase with the gate of Q2 due to the coupling transformer T1. This sets up a push-pull type circuit for the audio. It also causes nonlinear mixing or modulation across the FETs. The end result is that the mixing products or sidebands are in push-pull across the primary of T2 and, therefore, pass to the output. However, the carrier is canceled across the primary and thus, does not appear at the output. The original audio is filtered out by the RF transformer and resonant tank. Thus, the output is a double sideband, suppressed carrier signal.

The amount of carrier suppression depends upon how closely the FETs match and also the other components in the system. As mentioned before, R1 can be adjusted to improve the balance. Also, the center tap on the primary of T2 can be adjusted for optimum operation. When properly constructed and adjusted, this circuit can suppress the carrier to –60 dB below the sideband amplitude.

Diode-Ring Balanced Modulator

A balanced modulator that uses diodes as the nonlinear elements is shown in Figure 7-2. In this circuit, the diodes must be carefully matched to achieve a good carrier balance, or null.

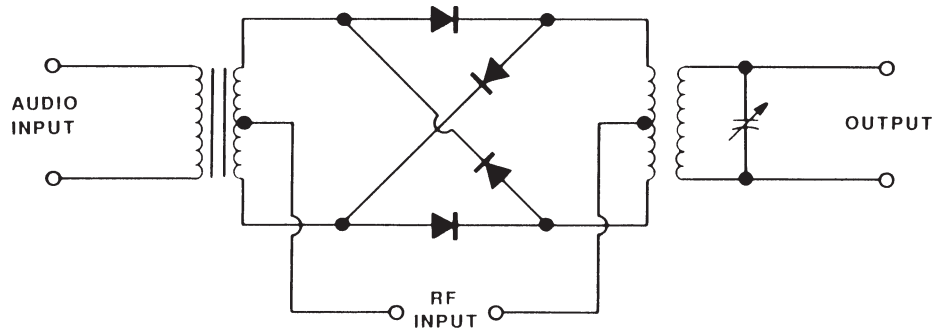


Figure 7-2

The diode-ring balanced modulator.

Figure 7-3A shows the circuit operation when the RF input has a polarity as shown. Diodes D1 and D2 are forward biased and equal but opposite currents flow in both side of T1. Therefore, there is no output. When the RF input polarity reverses, diodes D3 and D4 conduct as shown in Figure 7-3B. Here again, equal but opposite currents flow in T1 and there is no output.

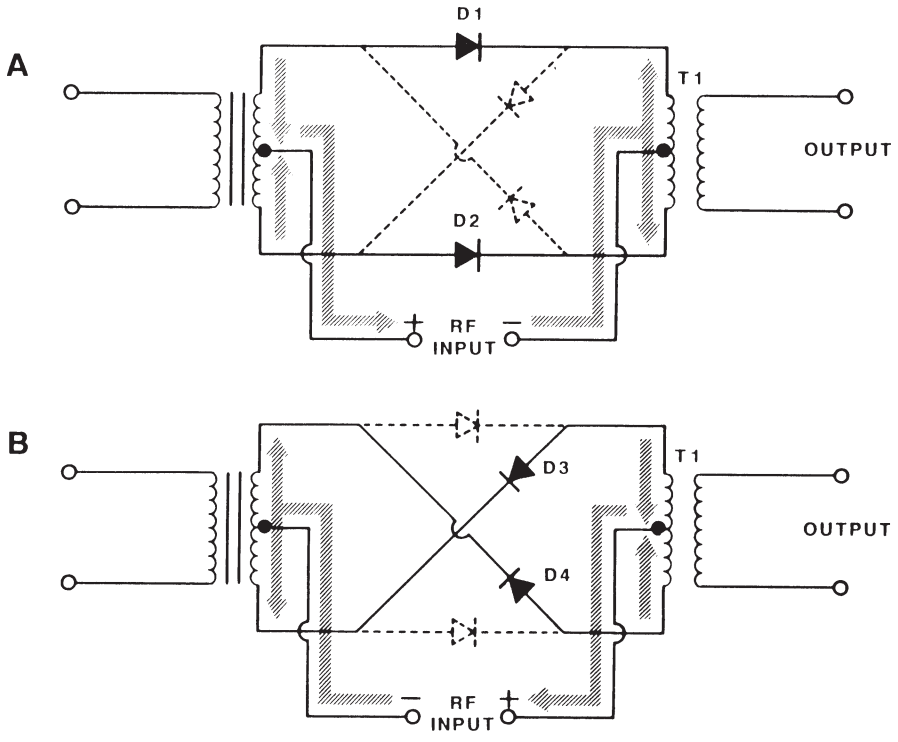


Figure 7-3

Diode-ring operation with only the RF carrier applied.

Figure 7-4A shows the circuit action when the audio signal is applied. To simplify the drawing, the RF input has been removed and the diodes replaced by switches. When diodes D1 and D2 conduct, due to the RF input, the audio signal flows through T1 in the direction shown. When the RF signal changes polarity, as shown in Figure 7-4B, diodes D3 and D4 conduct. This causes the signal through T1 to change polarity. Thus, the audio polarity at T1, is actually being switched back and forth by the RF input.

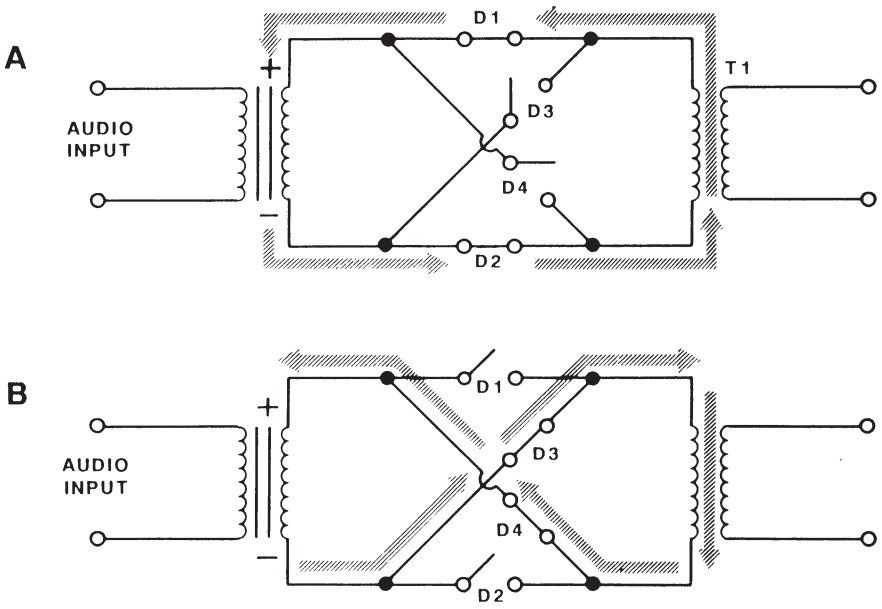


Figure 7-4

Audio signal switching action in the diode-ring modulator.

To clarify matters, the audio signal does not actually forward bias the diodes. Instead, the RF input switches the diodes and the audio signal sees the forward biased diodes as a low impedance path. This is due to the varactor action of the diodes. A forward biased diode has a high capacitance and, therefore, a low capacitive reactance to the audio signal. On the other hand, the reverse biased diodes offer a high reactance to the audio signal.

The various waveforms for this circuit are shown in Figure 7-5. The RF input is shown as a square wave. This allows instantaneous diode switching and thereby simplifies the drawing. Note that each time the RF input goes negative, the output wave reverses polarity. This is due to the diode switching action described earlier. You can follow this throughout the output waveform.

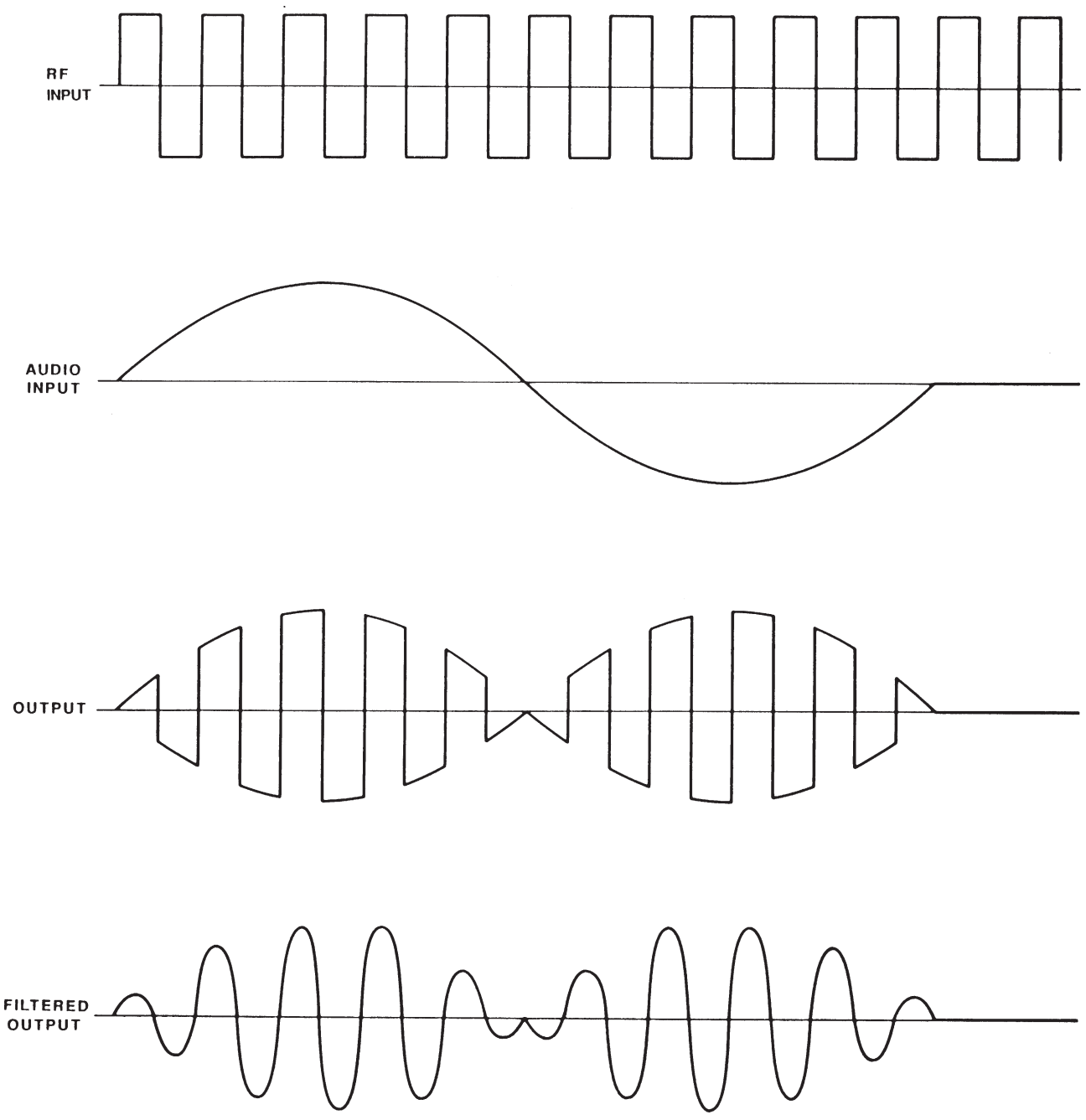


Figure 7-5

Waveforms in the diode-ring balanced modulator.

Notice that the output waveform's envelope is a "double" or "mirror image" of the modulating audio. This is significantly different than the standard AM envelope which duplicates the modulating signal. It can be shown that this complex wave is actually composed of two equal amplitude sine waves whose frequency separation is twice the modulating audio frequency. This corresponds exactly to upper and lower sidebands spaced around a suppressed carrier. In short, the output is a double sideband, suppressed carrier signal.

You might well ask: where did the sidebands come from when all the circuit did was switch diodes on and off? Since the diodes are being switched, they have a nonlinear resistance. Therefore, nonlinear mixing occurs, creating the mixing products and, thus, sidebands. The carrier is balanced out by circuit action.

The diode-ring modulator is widely used. In fact, it is often made in special packages using "hot-carrier" diodes. These devices offer wide operating frequency ranges and excellent carrier suppression.

There are several variations of the diode-ring balanced modulator. They all use diodes as the nonlinear elements and operate in a similar manner. That is, they balance out or null the carrier, and the nonlinear mixing across the diodes creates sidebands which are passed to the output.

IC Balanced Modulators

One circuit that lends itself extremely well to balanced modulator applications is the differential amplifier. A simplified diagram of a differential amplifier is shown in Figure 7-6. Q3 acts as the current source for Q1 and Q2. If the RF input is applied to the bases of Q1 and Q2 in phase, current through both transistors is identical and the voltage difference across the output is zero. The common-mode rejection characteristic of the differential amplifier has balanced out the carrier.

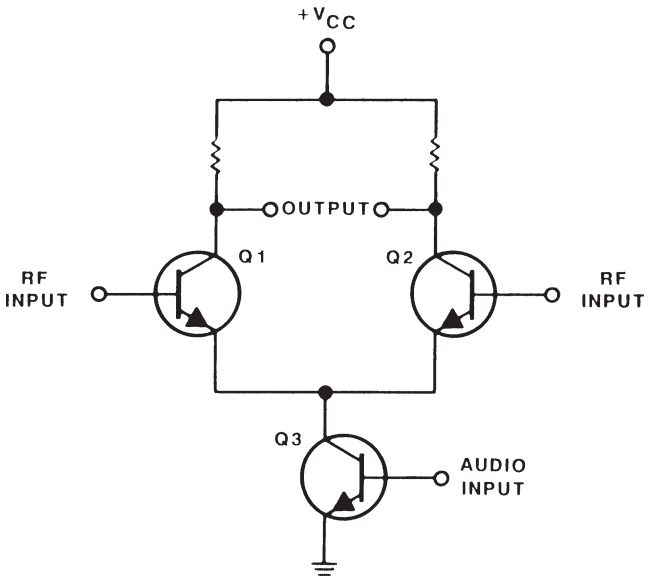


Figure 7-6

A differential amplifier used as a balanced modulator.

The audio input is applied to the base of Q3. This upsets the circuit balance. As a result, the audio and RF signals are mixed across Q1 and Q2. This is nonlinear mixing and, therefore, sidebands appear at the output. However, the carrier or RF input does not. Since it is a common-mode signal, it is rejected.

A differential amplifier must be constructed with transistors whose characteristics are very closely matched. Forming the transistors on a single chip of silicon, as is done with ICs, ensures this necessary match. Therefore, the differential amplifier is ideally suited to integrated circuit construction.

Figure 7-7 shows ICs that have been specially designed for use as balanced modulators. Figure 7-7A shows the 1496 balanced modulator. This device uses a differential amplifier configuration similar to what was previously described. Its carrier suppression is rated at a minimum of –50 dB with a typical value of –65 dB at 500 kHz.

Figure 7-7B is the SL1640 balanced modulator. It also uses a differential amplifier configuration. Its primary advantage is the small number of external components required. Its carrier suppression is typically –40 dB at 30 MHz.

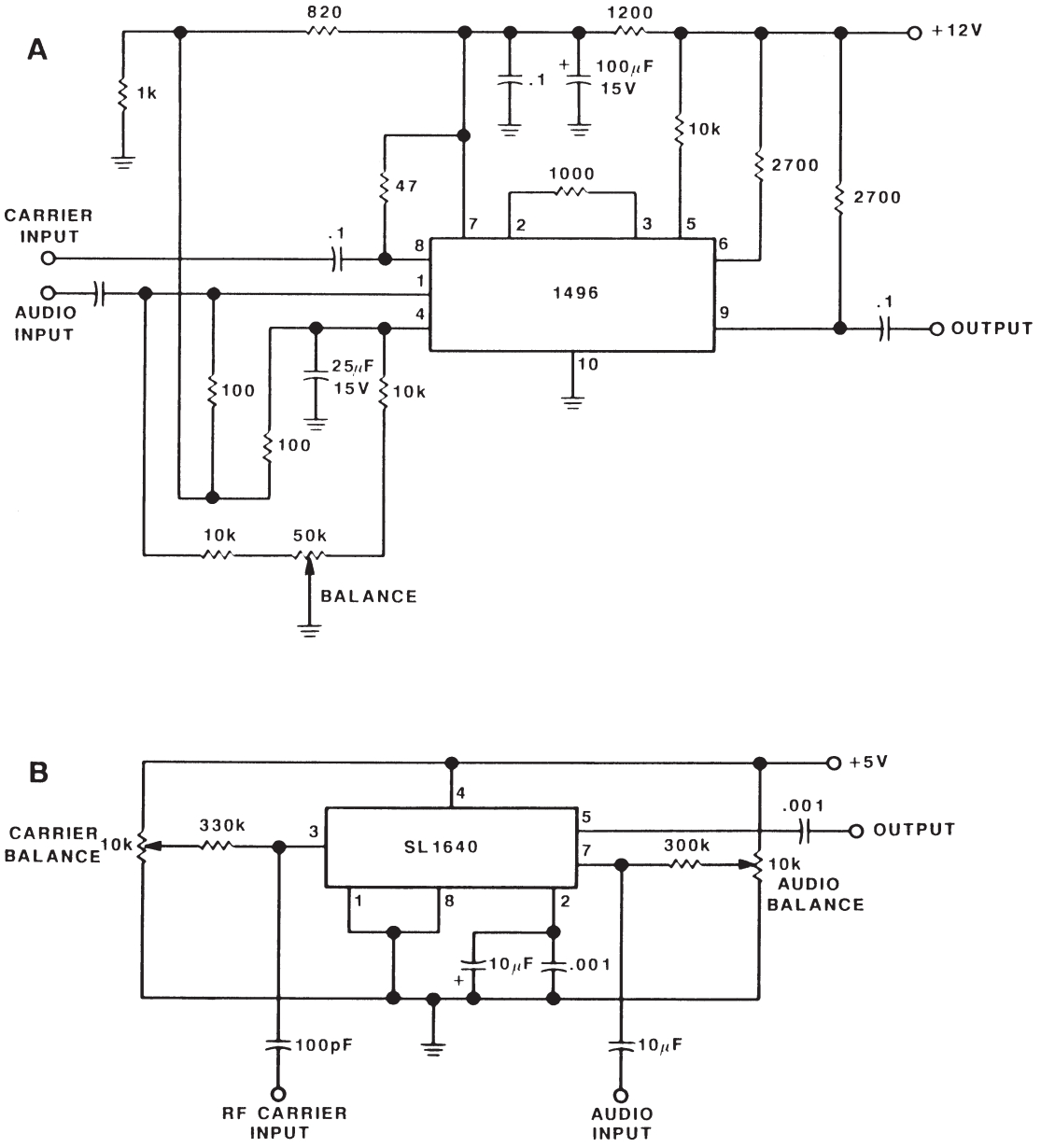


Figure 7-7

Integrated circuit balanced modulator.

DISADVANTAGES OF SUPPRESSED CARRIER AM

One of the disadvantages of standard AM is that most of the transmitted power is in the carrier. This power is wasted since it contains no intelligence. For this reason, the balanced modulator is used to suppress the carrier. This allows transmission of only the sidebands. However, the other disadvantages of AM remain: excessive bandwidth, propagation, and noise problems.

The bandwidth of a suppressed carrier AM signal is still twice that of the highest modulating signal. This is twice as much bandwidth as is needed to transmit the modulating information.

As mentioned before, for perfect AM reception, the two sidebands and carrier must be received exactly as they are transmitted. However, in suppressed carrier AM, the sidebands are transmitted without the carrier. This means that the carrier must be "reinserted" at the receiver in order to demodulate the signal. Since phase relationships are critical in AM, the reinserted carrier must not only be the same frequency as the original but also exactly in phase with the original. This is the most difficult problem to solve when designing a double sideband, suppressed carrier system. It can be solved, but unfortunately, the required circuitry is usually too complex and expensive for most practical applications. However, double sideband suppressed carrier signals are used as part of the color television system and the FM stereo broadcasting system.

A further disadvantage of suppressed carrier AM is that it is normally accomplished at low signal levels. This means that the modulated signal is generated at a low signal level, and must then be amplified to a level suitable for transmission. Since the modulated signal must be accurately preserved as it is amplified, these amplifiers usually operate at Class B or, more frequently, Class AB. Neither of these classes of amplifier is as efficient as the Class C amplifier that can be used to increase the carrier signal strength in standard AM, which is then modulated at a high level. Thus, some of the efficiency gained by suppressing the carrier is given up in the decreased efficiency of the transmitter amplifiers.

Trade-offs are involved in the choice of any system. Suppressed carrier AM provides increased signal efficiency at the cost of lowered transmitter efficiency and increased complexity. These characteristics limit the use of double sideband suppressed carrier AM to specialized applications.

SUMMARY

This chapter introduced you to suppressed carrier AM. As you learned in the previous chapter, half of transmitted signal power is wasted in the carrier. This results in only 25% of the power in each sideband. By reducing or eliminating the carrier, we can put more of the signal power in the sidebands.

A balanced modulator is used to suppress the carrier. The three most common types of modulators that are used are: the twin-FET balanced modulator, the diode-ring balanced modulator, and IC-balanced modulators.

Although suppressed carrier AM transmitters reduce or eliminate the power wasted in the carrier, the resulting signal still requires excessive bandwidth and is affected by propagation and noise. Furthermore, receivers need special circuitry to reinsert the carrier at the same frequency and phase as the original before the signal is usable. Another disadvantage is that suppressed carrier AM occurs at low signal levels and needs to be amplified.

SELF-TEST REVIEW

1. What are the disadvantages of standard AM?

   _____

   _____

2. Name three types of balanced modulators. _____

   _____

   _____

3. Identify the waveforms shown in Figure 7-8.

   A: _____

   B: _____

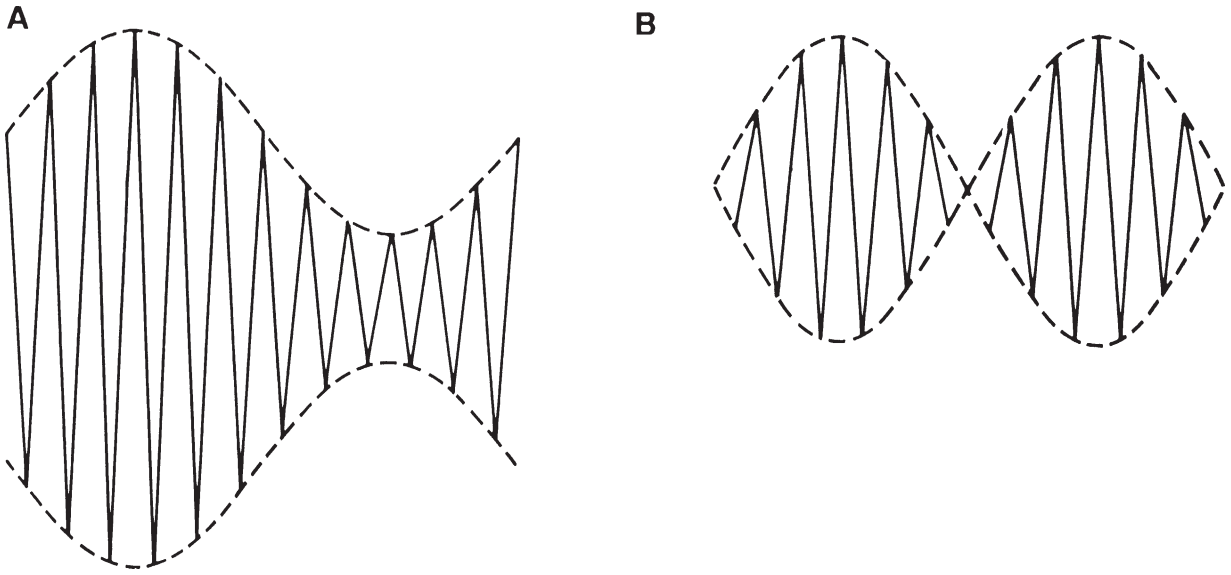


Figure 7-8
Figure for Question 3.

4. What is the primary advantage of suppressed carrier AM? What is the primary disadvantage?

_____

_____

5. What is the difference between low-level and high-level modulation? _____

_____

CHAPTER 8 — Single Sideband

INTRODUCTION

Most disadvantages of standard AM can be overcome by a system called single sideband suppressed carrier (SSB). In this system, only one sideband is transmitted. The carrier and the other sideband are suppressed at the transmitter. This system can still convey intelligence because either sideband contains all the information of the original modulating signal.

This approach requires a more complicated and, therefore, more expensive transmitter. The transmitter generally uses a balanced modulator to suppress the carrier and one of several methods to suppress the unwanted sideband. The design of the receiver is also complicated. As in double sideband suppressed carrier, a substitute carrier must be added to the received sideband. However, in SSB, the reinserted carrier must only be within a few Hertz of the original carrier. This is much easier to provide than the precise frequency and phase requirements for the reinserted carrier in double sideband suppressed carrier AM. As a result, SSB is less complex. Also, its many advantages over standard AM outweigh its additional cost and complexity.

The first advantage of single sideband is that all the power transmitted represents intelligence. Compare this to the standard AM system in which at least one-half of the transmitted power is in the carrier, and the remaining power is split between two sidebands, only one of which is needed. This means that a single sideband transmitter that is radiating 50 watts can produce the same audio signal level at the receiver as a standard AM transmitter that is radiating 200 watts.

Single sideband also reduces the bandwidth requirements. Since only one sideband is transmitted, the bandwidth is only one half that of a standard AM transmitter. Thus, a given frequency range can accommodate twice as many SSB channels as standard AM channels.

Finally, SSB is far superior to AM when propagation conditions are poor. SSB can be up to eight times more effective in getting through when propagation conditions are really bad. This is due primarily to its narrow bandwidth and the power advantages discussed above.

OBJECTIVES

When you complete this chapter, you will be able to:

- Discuss the advantages of single sideband over standard AM and suppressed carrier AM.
- Describe two methods of generating an SSB signal.
- Describe the operation of two types of sideband filters.
- Draw a block diagram of a filter type SSB transmitter and show the frequencies present at each stage.
- List three methods of sideband selection.
- Compute the peak envelope power rating of an SSB transmitter.
- Compare and contrast the practical use of AM, suppressed-carrier double sideband, and single sideband.

FILTER METHOD SSB TRANSMITTER

The most widely used technique for generating a single sideband signal is the filter method. A block diagram of this type of transmitter is shown in Figure 8-1. Here, the modulating audio is amplified and, along with the RF carrier, applied to the balanced modulator. The output of the balanced modulator, which is a double sideband suppressed carrier signal, is then applied to the sideband filter. This filter is designed to pass the desired sideband and block the unwanted sideband. The output is either an upper sideband signal or a lower sideband signal, depending on the filter's passband.

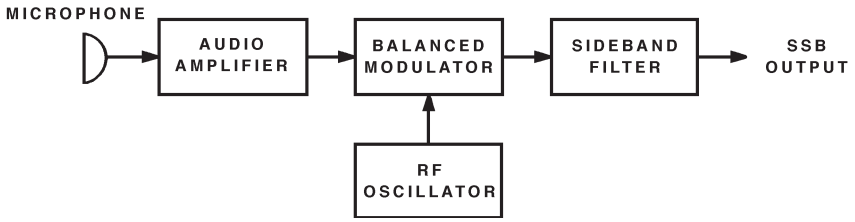


Figure 8-1
Filter method SSB transmitter.

Sideband Filters

In a single sideband transmitter, the input audio is usually limited to the 250 Hz to 3000 Hz range. This is all that is required for adequate speech intelligibility and it narrows the signal bandwidth still further. It is accomplished by using a simple audio filter in the input audio amplifier. The lower limit of 250 Hz results in a 500 Hz separation between the upper and lower sidebands as shown in Figure 8-2A.

The function of the sideband filter is to fully attenuate one sideband while allowing the other sideband to pass. Therefore, the filter must drop from 0 dB attenuation to approximately –50 dB attenuation within the 500 Hz frequency range that separates the sidebands. This requires a very sharp selectivity "skirt" or drop from 0 dB to –50 dB. A typical sideband filter selectivity, or response, curve is shown in Figure 8-2B. In this case, the filter allows the upper sideband to pass and attenuates the lower sideband.

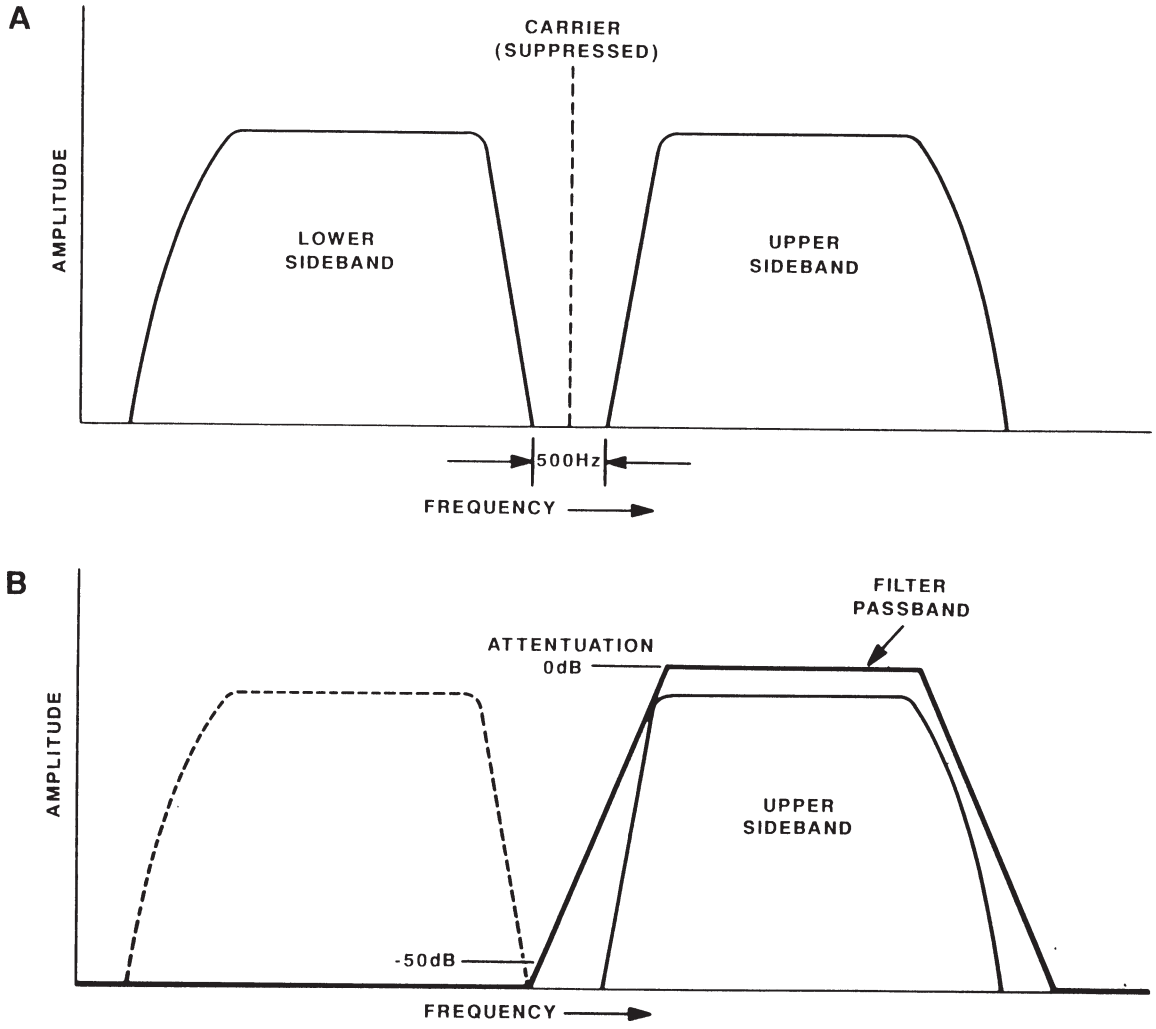


Figure 8-2

SSB filter passband.

One of the most widely used sideband filters uses quartz crystals to achieve the desired passband shape. Let's briefly review the characteristics of crystals before continuing.

Figure 8-3A shows the equivalent electrical network for a quartz crystal. The inductance L represents the mass of the crystal, C_1 represents the piezoelectric resilience of the crystal, R represents the frictional losses, and C_2 is the capacitance of the metal electrodes holding the crystal. In most crystals, the frictional losses (R) are extremely small and, therefore, the circuit Q is very high. In fact, the Q ranges from 10,000 up to 1 or 2 million. This makes crystals ideally suited for use in filters.

The impedance curve for a quartz crystal is shown in Figure 8-3B. This shows that at low frequencies, the network impedance is capacitive and high. As frequency increases, the impedance drops until it reaches f_s. This is the series resonant frequency of the crystal which is determined by the crystal's cut and thickness. Going back to the equivalent circuit, the reactances of L and C, are equal, at f_s, forming a series resonant circuit. At a slightly higher frequency, f_p, the inductive reactance of L is equal to X_{c2}. Therefore, at this frequency, the crystal is a parallel resonant circuit. These characteristics of the quartz crystal call be used to form a sideband filter.

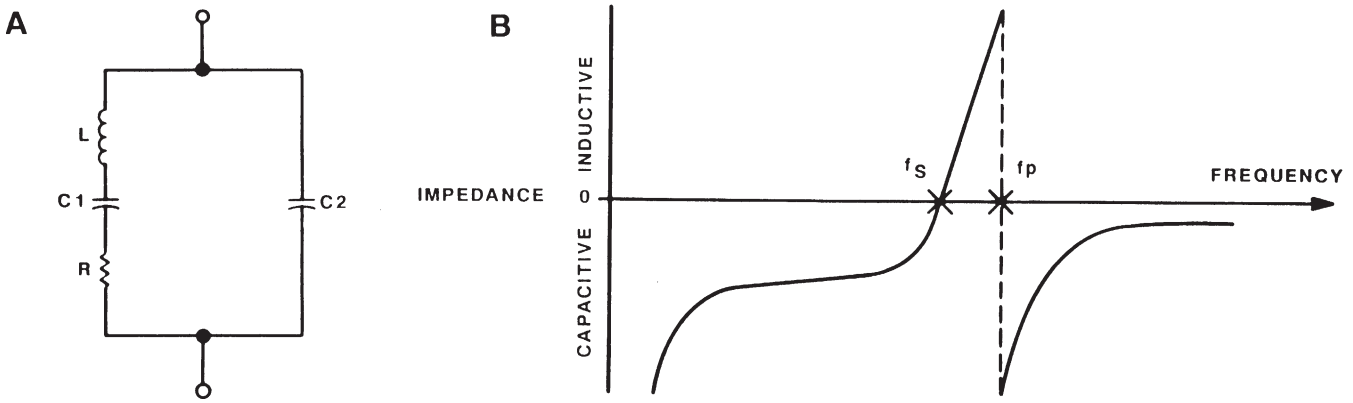


Figure 8-3

(A) Equivalent circuit of a quartz crystal.
(B) Impedance of the crystal versus frequency.

A crystal lattice filter is shown in Figure 8-4. This filter uses two matched pairs of crystals in a special bridge network. In this circuit, X_{1A} and X_{1B} are identical to each other, and X_{2A} and X_{2B} are also identical to each other but slightly different from X_{1A} and X_{1B}. The impedance diagram for the crystals is shown in Figure 8-5. Note that there are two parallel resonant frequencies for each crystal. This is due to the inductance present in the circuit. They will form resonant circuits with the crystals at specific frequencies. Let's analyze this circuit in detail.

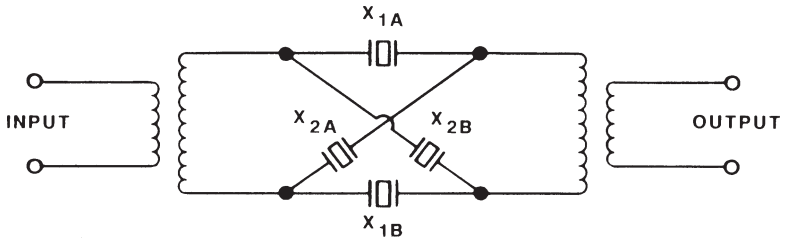


Figure 8-4

A crystal lattice filter.

At f_1 in Figure 8-5, both X_1 and X_2 present equal impedances to the input signal. Therefore, in the circuit of Figure 8-4, equal amplitude signals are passed through X_{1A} and X_{2A}. However, since they are from opposite sides of the transformer, they are 180° out of phase, and cancel. The same holds true for X_{1B} and X_{2B}. This is a point of maximum attenuation for the filter as shown on the response curve of Figure 8-5.

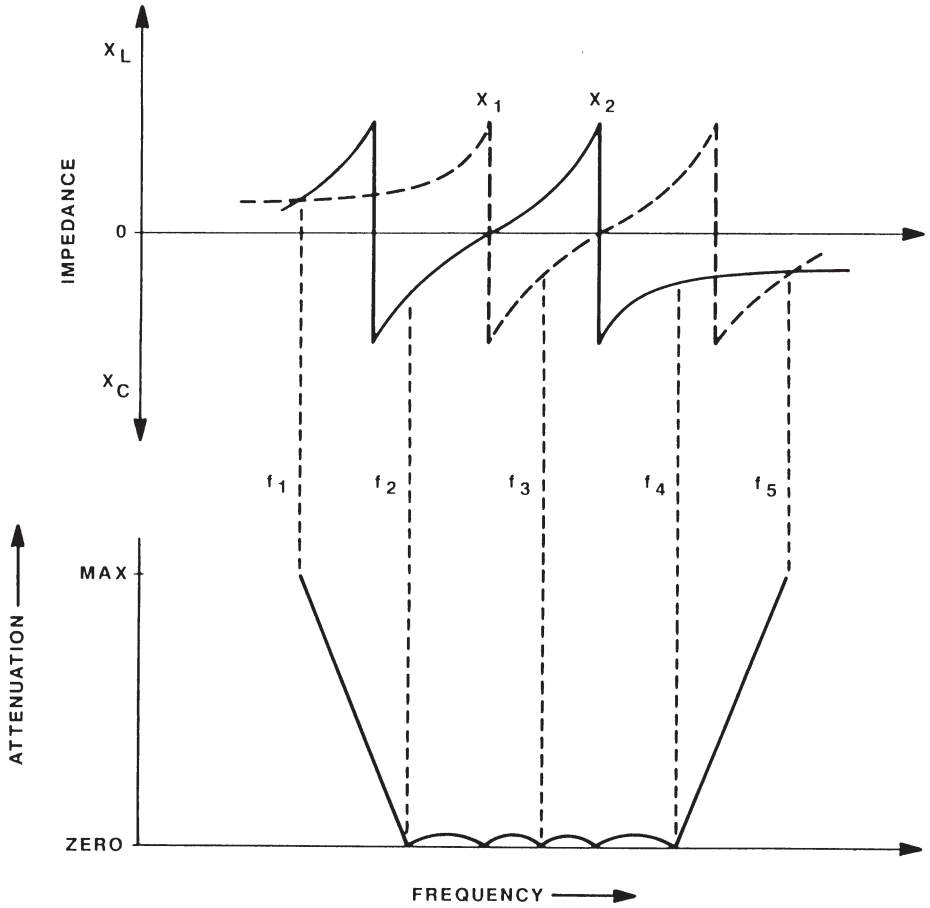


Figure 8-5

Impedance diagram and selectivity curve of a crystal lattice filter.

At f_2, both X_1 and X_2 have equal impedances again. However, X_1 acts inductively while X_2 acts capacitively. The result is that the signals arrive at the output in phase. For example, if the signal at X_{1A} is at 0°, since X_{1A} acts inductively, it will shift the signal +90°. The signal at X_{2A} is at +180° due to the transformer. X_{2A} acts capacitively and, therefore, shifts the signal –90°. The result is that both signals arrive at the output with a phase angle of +90°. They are in phase and there will be zero attenuation at this frequency. This is shown in Figure 8-5.

At a frequency midway between f_2 and f_3, X_2 attains series resonance while X_1 is in parallel resonance. The signal, therefore, passes unattenuated to the output through X_{2A} and X_{2B}. At f_3, the crystal's impedances are again equal but opposite and zero attenuation occurs. Midway between f_3 and f_4 crystal X_1 reaches series resonance and yet another zero attenuation point occurs. At f_4 again, the impedances are equal but opposite and zero attenuation occurs. However, at f_5, the impedances are equal and in phase. The output signals, therefore, cancel and maximum attenuation occurs.

As you can see, the crystal lattice filter makes a good sideband filter with relatively steep attenuation curves, or "skirts," On both sides of the passband. However, for most applications, even steeper skirts and more attenuation of the unwanted sideband are required. For this reason, several lattice sections must be cascaded to achieve the required response.

Another type of filter used extensively in single sideband transmitters is the mechanical filter. A typical mechanical filter is shown in Figure 8-6. The coupling coil along with the biasing magnet form an electrical-to-mechanical, or magneto-striction, transducer. Any electrical signal across the coupling coil sets up a magnetic field which either opposes or attracts the biasing magnet. This magnetic action causes the coupling coil and transducer rod to move. This mechanical motion is then coupled to the disc resonators. These discs are cut to precise dimensions and, therefore, resonate at only a certain band of frequencies. When the discs resonate, they pass the signal on to the output transducer. If they do not resonate, they attenuate the input signal substantially.

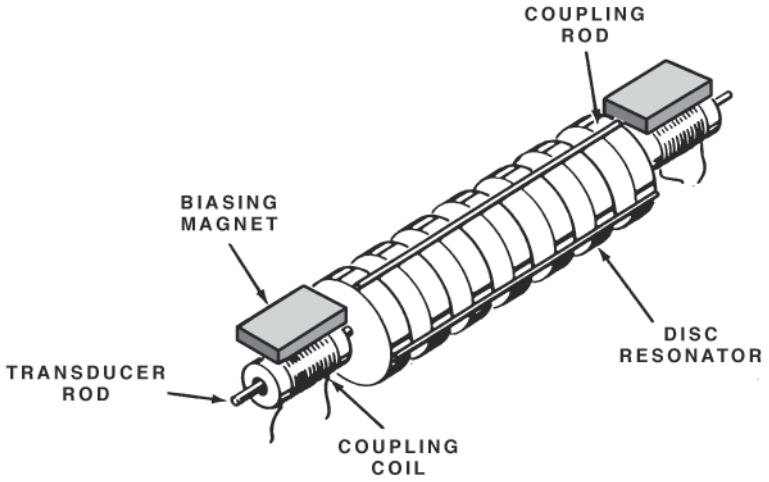


Figure 8-6

A sideband filter using mechanical disc resonators.

Mechanical filters offer excellent characteristics and are often chosen for the most exacting applications. However, they are limited to operation in the 60 Hz to 600 kHz range, whereas crystal filters can be constructed for signals up to 10 MHz and higher.

Frequency Conversion

Thus far in the SSB transmitter, we have an upper, or lower, sideband signal present at the filter's output. However, this is a fixed frequency signal. That is, the sideband filter is fixed-tuned and, therefore, the SSB signal is also restricted to a single frequency. We must now convert this SSB signal to the desired operating frequency.

Figure 8-7 shows a typical frequency conversion technique. It uses a mixer and a heterodyne oscillator. The mixer is any nonlinear circuit and could even be a balanced modulator. The heterodyne oscillator is a crystal oscillator, a variable frequency oscillator, or a combination of both. Let's examine Figure 8-7 in detail.

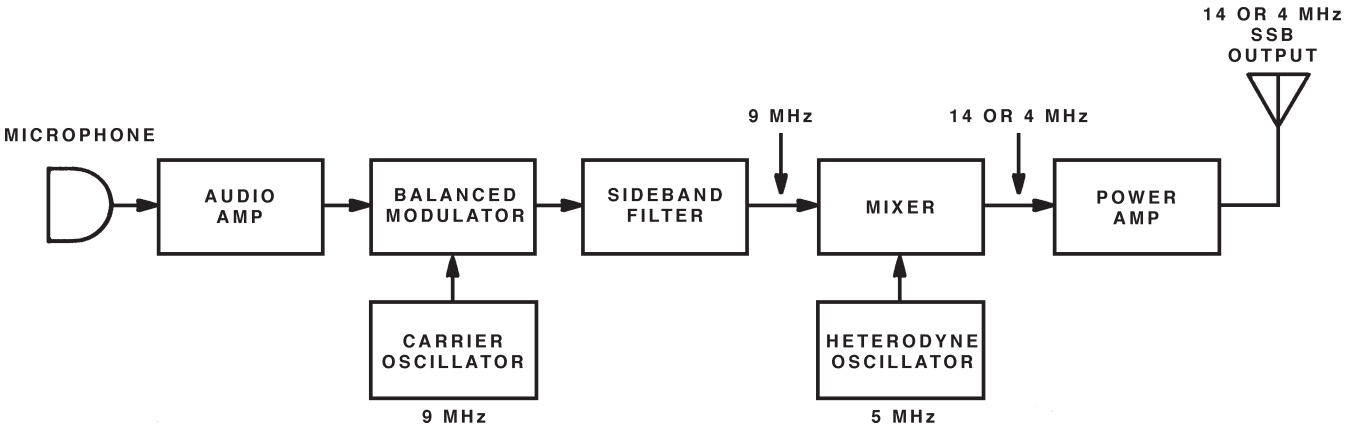


Figure 8-7
Complete SSB filter transmitter.

The carrier oscillator is crystal controlled at 9 MHz, which is the passband of the sideband filter. Therefore, the SSB signal, either upper or lower sideband, is applied to the mixer at an approximate frequency of 9 MHz. The heterodyne oscillator frequency for this example is 5 MHz, although it may be switched or varied to obtain any desired output frequency. In this case, the 9 MHz and 5 MHz signals are applied to the mixer, and the result is sum and difference mixing products. Therefore, either 4 MHz or 14 MHz appears at the output. The desired mixing product is selected by a resonant tank circuit at the output of the mixer. The power amplifier is also tuned to the desired frequency. It amplifies the low-level SSB signal to the needed level for transmission. Thus, once the SSB signal is generated at a low-level fixed frequency, it is converted to the desired operating frequency using a heterodyne oscillator and a mixer.

Sideband Selection

There are three methods of selecting either the upper or the lower sideband. The obvious method, using two different filters, is shown in Figure 8-8A. The only drawback to this method is the additional cost of the second filter. The most commonly used technique is shown in Figure 8-8B. This method shifts the frequency of the carrier oscillator. For upper sideband, the carrier oscillator frequency is set to the lower edge of the filter's passband. For lower sideband, it is set to the upper edge of the filter's passband. The carrier frequency is usually changed by switching between two crystals.

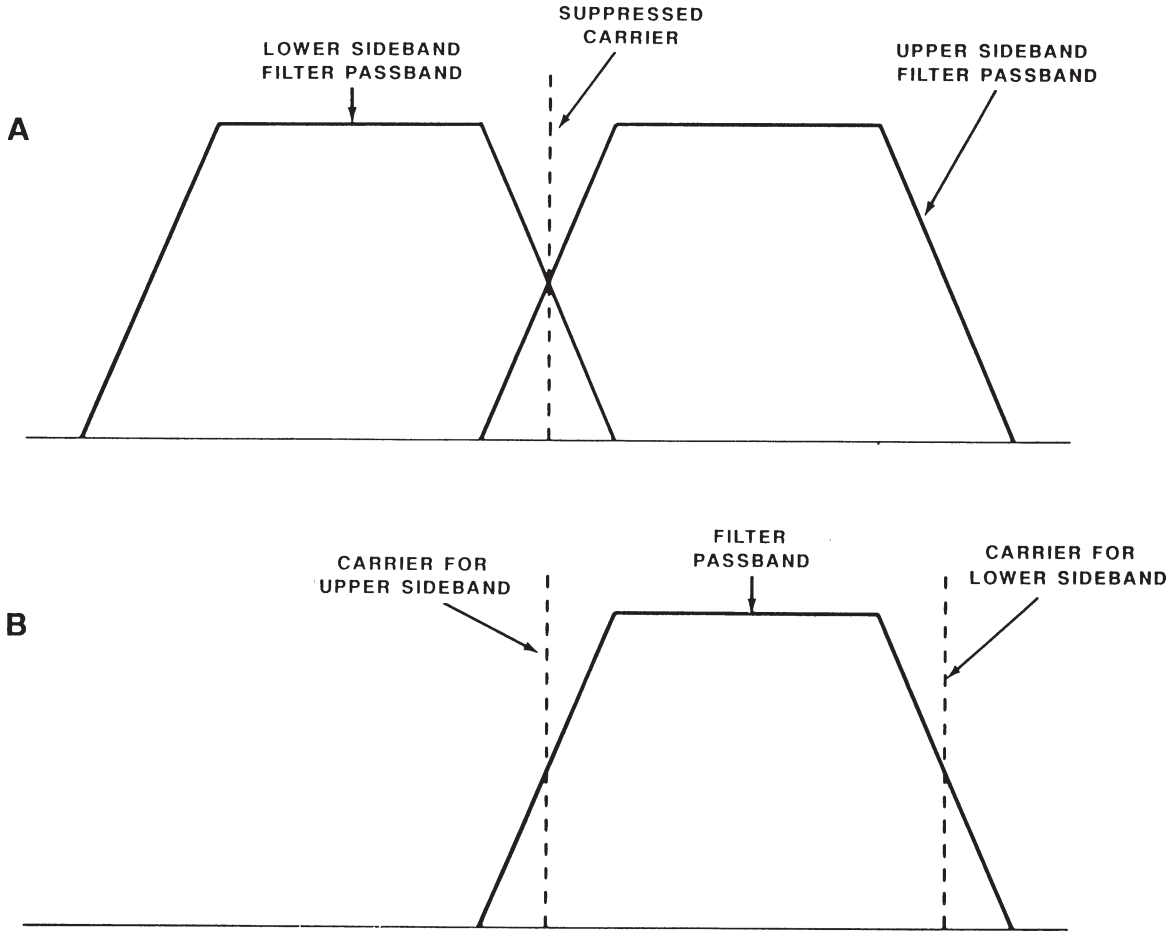


Figure 8-8

Two methods of sideband selection. Phasing method SSB transmitter.

The third method of sideband selection involves frequency inversion. This can occur during the heterodyne or frequency conversion process. As an example, suppose an upper sideband signal is generated with suppressed carrier of 9 MHz and sideband frequencies extending to 9.003 MHz. When this signal is mixed with a 15 MHz heterodyne oscillator signal, the sum products will be a 24 MHz suppressed carrier with sideband components out to 24.003 MHz. This is still an upper sideband signal. However, the difference products will be a 6 MHz suppressed carrier and sideband components down to 5.997 MHz. This is now a lower sideband signal.

Frequency inversion can be used to select the desired sideband or one of the other two methods. However, whichever method is chosen, the effects of frequency inversion must be taken into account when designing the transmitter.

PHASING METHOD OF SSB

The major disadvantage of the filter method is the expensive sideband filter required. To get around this, the phasing method of generating a single sideband signal was developed. The block diagram for this system is shown in Figure 8-9.

With this system, a 90° phase shift network is used to split the audio signal into two components that have a 90° phase difference. The RF oscillator signal is also split into two components that have a 90° phase difference. One RF and one audio signal are then applied to each of the two balanced modulators.

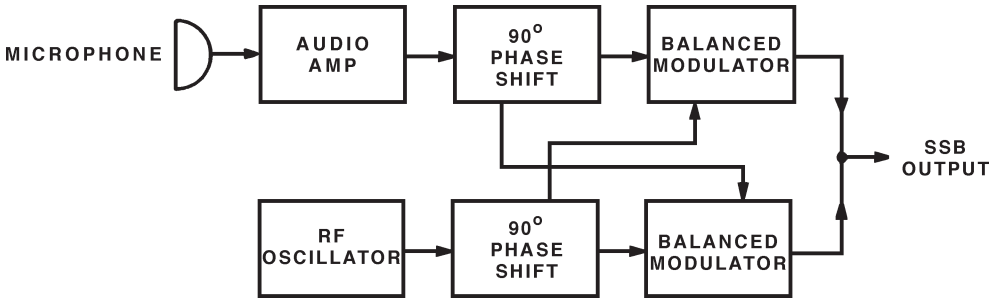


Figure 8-9

Phasing method SSB transmitter.

Both balanced modulators suppress the carrier and have a double sideband output. These output signals are then added together and, due to the phasing of the applied signals, one sideband is canceled and the other sideband is reinforced.

The disadvantage of the phasing method is that it requires frequent and precise adjustment to operate correctly. For this reason, the filter method is commonly used in modern single sideband transmitters.

LINEAR AMPLIFIERS

A single sideband signal is usually generated at a very low power level, often less than one watt. To increase its power level, one or more power amplifiers must be used. These power amplifiers must be linear; they must amplify the SSB signal without introducing any distortion. Otherwise harmonics and other forms of "spurious radiation" will appear at the output and cause interference.

For linear amplification of an SSB signal, the amplifier must be biased for class A, AB, or B operation. Class AB and B can be used because the flywheel effect of the output tank completes the RF waveform. However, class C amplifiers cause serious distortion and cannot be used. Normally class AB or B amplifiers are used because they offer the best efficiency.

SSB POWER RATINGS

The power rating for an AM transmitter is its DC input power under "carrier only" conditions, or in other words, with no modulating signal. However, a single sideband transmitter does not have a carrier; with no modulating signal, there is no output. This makes power measurement slightly more difficult with an SSB transmitter.

Figure 8-10 shows the output waveform of an SSB transmitter with a normal speech modulating signal. The power contained in the maximum peak of the signal is the basic transmitter rating and is called the peak envelope power (PEP). However, envelope peaks occur only sporadically during voice transmission. Therefore, an average power reading is much easier to obtain. In fact, the final amplifier current and voltage meters respond only to the average value. This is because the needles of the meters are unable to respond to the rapid peaks and thus give an average indication.

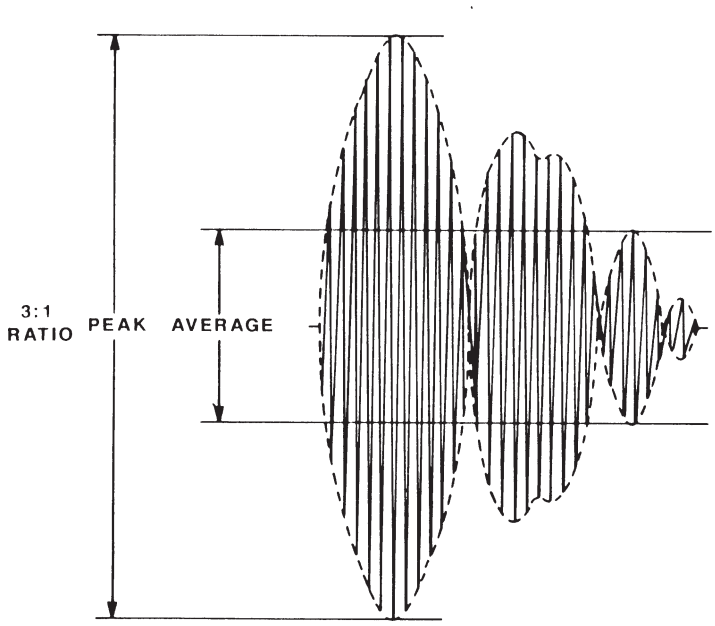


Figure 8-10
SSB output waveform.

To determine the PEP input from the average meter readings, you must determine the ratio between peak and average. This can be very difficult to do in the case of voice modulation because each person's voice characteristics are different. However, a two-tone test signal may be used that consists of two audio sine waves with a frequency separation of approximately 1,000 Hz. This test signal is then applied to the input of the SSB transmitter and the resulting output waveform appears as shown in Figure 8-11.

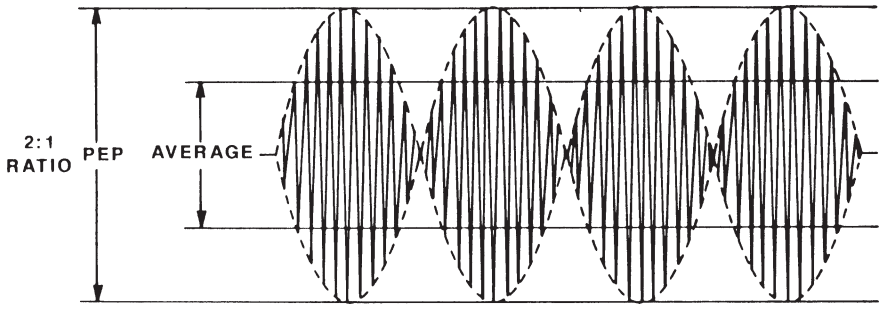


Figure 8-11
Two-tone SSB output.

With a two-tone test signal, the ratio between PEP input and average power input is 2:1. Therefore, the PEP rating is found by multiplying the input power reading (I × E) by 2. The ratio for a voice-modulating signal may be anywhere in the range of 1.5:1 to 7:1. Therefore, the two-tone test signal gives an accurate indication of the transmitter's PEP rating.

Because of its improved efficiency and reduced bandwidth, SSB is used in a wide variety of applications ranging from radio broadcast to wired telephone systems.

SUMMARY

In this chapter you learned about single sideband. Although single sideband transmitters and receivers are more complex and expensive to make than they are for standard AM, their advantages outweigh the cost and complexity.

As you learned earlier, half of the power of a standard AM transmitter's signal is wasted in the carrier while the other half is divided between the two sidebands. Double sideband suppressed carrier reduces the power wasted in the carrier, so that most of the power is used by the

sidebands. Single sideband goes one step further by removing the carrier and one sideband. Only one sideband is actually required, because the intelligence contained in each sideband is identical. Therefore, a single sideband transmitter needs to radiate only 1/4 the power that a standard AM transmitter would to result in the same audio signal level.

Single sideband transmitters usually contain a balanced modulator to remove the carrier and either a crystal lattice or mechanical disc resonator filter to narrow the signal's bandpass. A mixer circuit then mathematically combines the output of a heterodyne oscillator with the output of the sideband filter to convert the signal to the desired operating frequency. We then have to remove the unwanted sideband. This can be done by using a bandpass filter, shifting the frequency of the carrier oscillator to either the upper or lower edge of the filter's passband, or using frequency inversion.

Another less popular way of generating an SSB signal is the phasing method. This method works by applying phase shifted RF and audio signals to two separate balanced modulators and then adding them together. This cancels one sideband and reinforces the other.

Peak envelope power is used to describe the output power of an SSB transmitter. To do this, a two-tone audio test signal is applied to the transmitter and then the input power reading is multiplied by 2.

SELF-TEST REVIEW

1. What are the advantages of SSB over standard AM and double sideband suppressed carrier AM? _____

   _____

2. Name two methods of generating an SSB signal. _____

   _____

3. What two types of sideband filters are used in a filter method SSB transmitter? _____

   _____

4. What are the advantages and disadvantages of the two sideband filters mentioned in Question 3? _____

   _____

5. Label each block in the diagram of a filter method, single sideband transmitter shown in Figure 8-12.

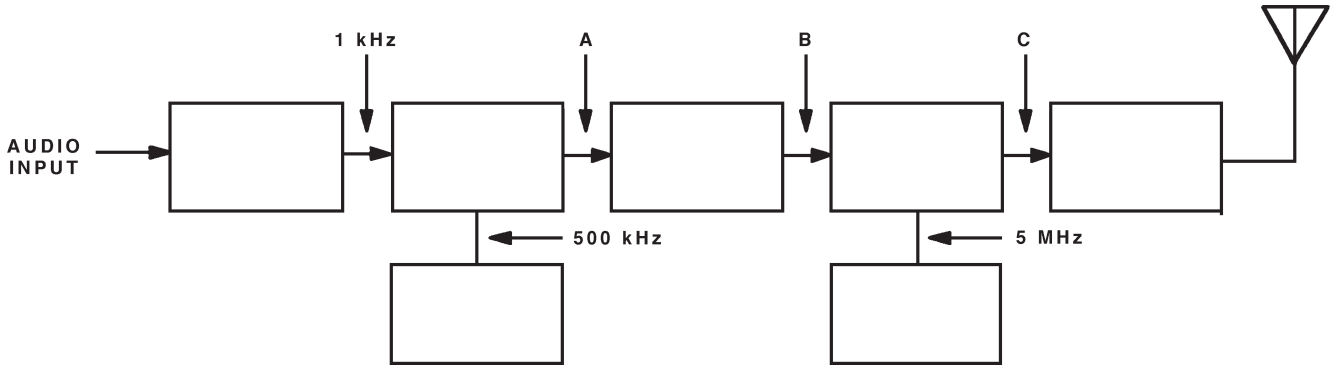


Figure 8-12

Figure for Questions 19 and 20.

6. In Figure 8-12, if the input is a 1 kHz sine wave with the other signals at the frequencies shown, what frequencies will be present at point A? _____

_____

What frequency will be present at point B for an upper sideband output? _____

_____

What frequency will be present at point C if the sum mixing product is used? _____

_____

What frequency will be present at point C if the difference mixing product is used? _____

_____

If the difference mixing product is used, will the output be an upper or lower sideband signal? _____

_____

7. List the three methods of sideband selection. _____

_____

8. With a two-tone modulating signal, the final amplifier ammeter indicates 3.4 A and the voltmeter indicates 28 V. What is the DC input power? _____

   What is the peak envelope input power (PEP)? _____

Section 3

AM Receivers

CHAPTER 9 — AM Detectors

INTRODUCTION

At the beginning of electrical communications, AM was used in radio. The first radio, or "wireless," receivers consisted of only a detector and a headset, for there was rarely more than a single station that could be heard. However, as more and more transmitters began to operate, this simple receiver was swamped with interference.

In the previous chapters you learned about amplitude modulation and the circuitry necessary to generate it. Beginning with this chapter you will study the circuits and techniques used to receive and detect amplitude modulated signals. Because you are familiar with radio receivers, and can easily relate the lesson to your experience, many of the examples used in these chapters involve radio receivers. However, AM signals are also found in many forms of communications which do not involve radio. You should study these chapters with the view that AM is not limited to radio and that the examples presented have a broad application.

As you learned, in the transmitter the carrier is modulated by causing amplitude variations in the transmitted signal that correspond to the amplitude variations in the modulating signal. By one means or another, this signal is propagated from the transmitting location to someplace else, where the signal strength is usually only a tiny fraction of what was transmitted. In addition to the desired signal, there is probably a number of undesired signals available at this receiving location, too. The receiver must first select the desired signal, and must then amplify this weak signal while it rejects all others. Finally, to recover the information that was sent, the receiver must demodulate the signal to recover the original modulating waveform. The circuit that does this is called a demodulator, or a detector.

OBJECTIVES

When you complete this chapter, you will be able to:

- List the four types of AM detectors and state the characteristics of each.
- Describe the requirements of the local oscillator signal to demodulate a double-sideband, suppressed-carrier signal.
- List the primary applications of double-sideband, suppressed-carrier AM.
- Describe the type of demodulator used for single-sideband signals.
- Describe the requirements of the local oscillator signal to demodulate a single-sideband signal.

THE DIODE DETECTOR

The most popular AM demodulator is the diode detector. This circuit is very simple and is used in virtually all AM receivers. Its purpose is to recover the modulation envelope from the AM waveform.

The diode detector is shown in Figure 9-1. Switch S1 is included merely for explanation. The input to the circuit is an AM waveform that has been selected and amplified by previous stages in the receiver. It is applied to diode D1, which acts as a half-wave rectifier. The positive half cycles cause D1 to conduct, developing positive pulses across R1. D1 cuts off the negative half cycles of the RF input. The center waveform shows the voltage developed across R1 if S1 is open.

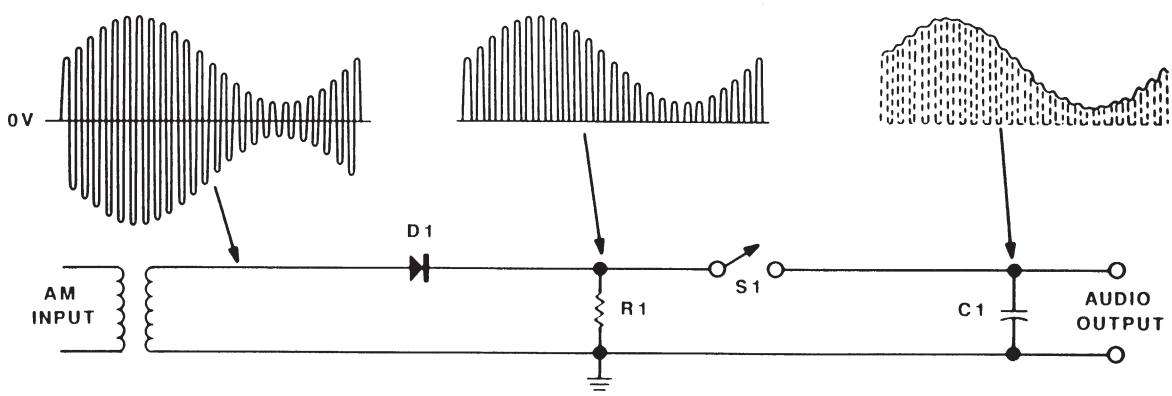


Figure 9-1

The diode detector.

When S1 is closed, C1 is placed in parallel with R1. C1 quickly charges through D1 to the peak of each positive pulse. Between pulses, C1 attempts to discharge through R1. However, the RC time constant is chosen so that C1 discharges only slightly. The result is that the voltage across C1 follows the envelope of the AM waveform. Thus, the output looks like the upper envelope with a small amount of ripple. Normally, the carrier frequency is many times higher than the envelope frequency, and therefore, the ripple is not noticeable.

Figure 9-1 shows the diode detector's operation in the time domain. Let's analyze its operation in the frequency domain.

The AM input consists of three frequency components: the carrier, the upper sideband, and the lower sideband. These signals are applied to D1 and are mixed across its nonlinear resistance. The difference signal is the modulating information. The next step is to separate this low frequency signal from the high frequency carrier. This is accomplished by C1 which acts as a short circuit to ground for the carrier signals and a high reactance for the audio signals.

The final step in AM detection is to separate the audio from the DC component. This is done quite simply using a coupling capacitor such as C2 in Figure 9-2.

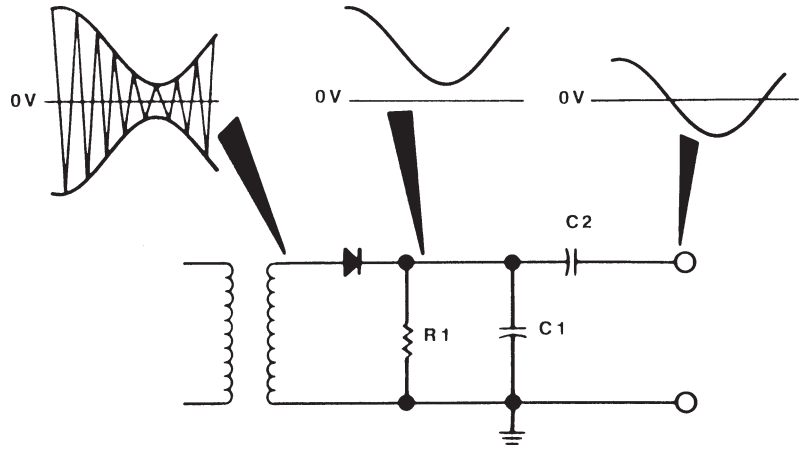


Figure 9-2
Complete diode detector circuit.

THE TRANSISTOR DETECTOR

Many low cost transistor receivers use a diode as the detector. However, due to the limited gain usually preceding the detector, a transistor is sometimes used as the demodulator to provide additional gain. A transistor can perform as a detector if it is biased for class B operation. In this way, the AM signal is both amplified and rectified at the same time.

Figure 9-3 shows a typical transistor detector. Resistors R1 and R2 form a biasing network which sets the circuit for class B operation.* R3 is the collector load resistor, while C1 filters out carrier components. This leaves the audio signal to be coupled to the output by C2.

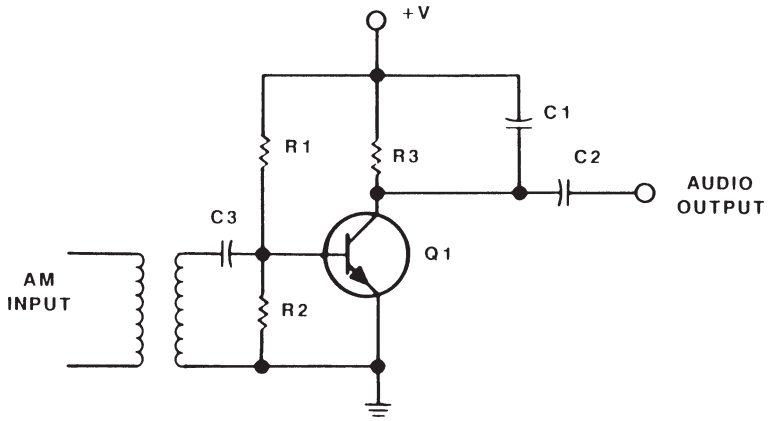


Figure 9-3

The transistor detector.

The transistor detector offers a method of achieving additional receiver gain. However, in anything but very low cost receivers, the diode detector is chosen for its simplicity and excellent performance.

* Note: A class B amplifier conducts only on positive input cycles, and cuts off when the input swings negative. So, it acts like a rectifier with gain.

THE SYNCHRONOUS DETECTOR

The diode and transistor detectors recovered the modulation envelope from the carrier signal by rectification. However, with double-sideband, suppressed-carrier AM, the envelope does not duplicate the modulating signal. This is shown in Figure 9-4 for a 10 kHz modulating tone and a 100 kHz carrier. Due to the absence of the carrier, the two sidebands combine to produce the envelope shown in Figure 9-4D. This envelope actually has a frequency of twice the modulating frequency or 20 kHz. You can see that this signal could not possibly be demodulated using a diode detector. The recovered signal would have severe distortion and be at twice the frequency of the original modulating information. Therefore, a special circuit called **a synchronous detector** must be used to demodulate a suppressed-carrier AM signal.

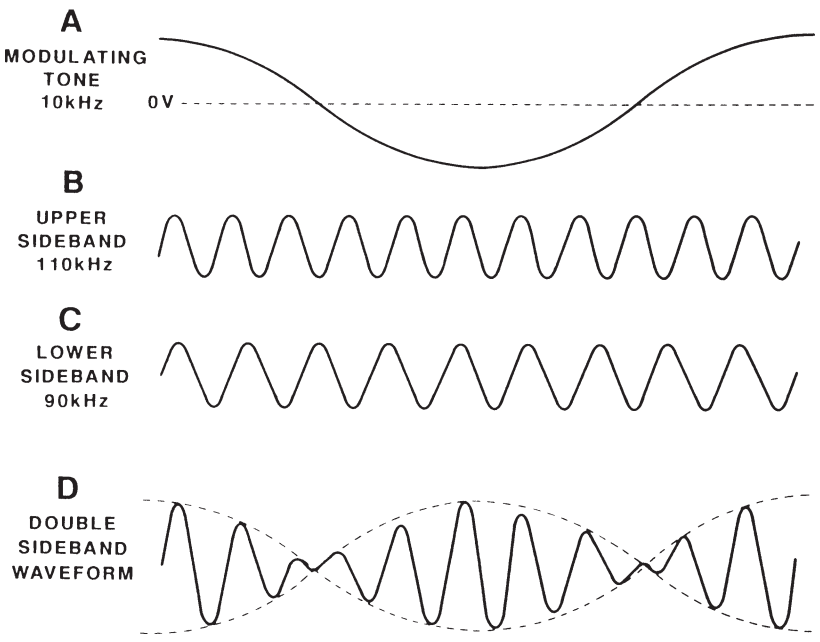


Figure 9-4
Moving the carrier.

A typical synchronous detector is shown in Figure 9-5. It consists of two diodes rather than the single diode used in the simple detector. It also has a separate "local carrier" input. This is a signal that is generated by the receiver. As you will see later, this local carrier signal must not only be the same frequency as the original suppressed carrier but must also be exactly in phase with the original.

In Figure 9-5, the local carrier input causes diodes D1 and D2 to switch. When the carrier input results in a negative potential at the top of T1 and a positive potential at the bottom, both D1 and D2 conduct. However, no current flows to the output or C4 because all current remains in the D1, D2, and T1 loop. In effect, the carrier input is balanced and, therefore does not reach the output. On the other half-cycle of the carrier, D1 and D2 are reverse biased and no current flows. The combinations of R1/C1 and R2/C2 form diode loads and ensure that the diodes are switched on for only a small fraction of each cycle.

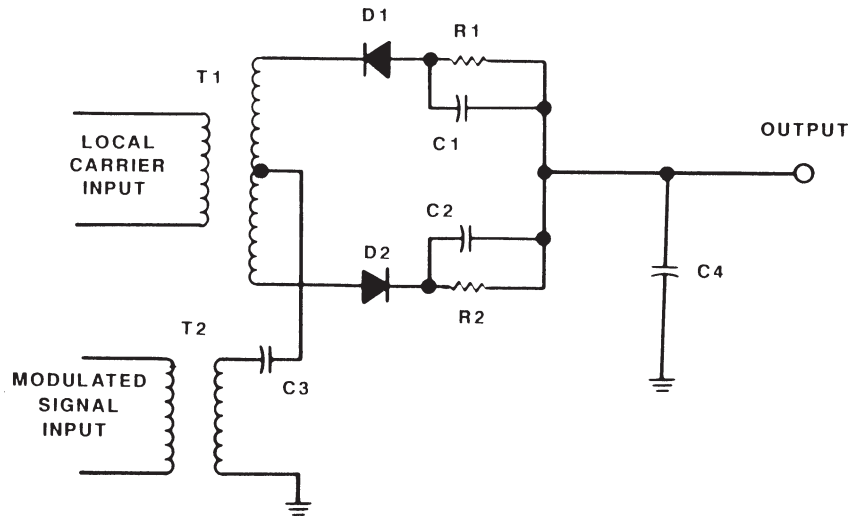


Figure 9-5

A synchronous detector.

The modulated signal input is of a sufficiently low level that it cannot forward bias D1 and D2. This means that the only time the modulated signal can reach C4, and the output, is when the local carrier input forward biases the diodes. What actually happens is that the carrier input causes samples to be taken of small portions of the modulated signal as shown in Figure 9-6. Here, every time the carrier input goes positive, the diodes are turned on and the suppressed carrier signal passes to the output. Note that the carrier samples the positive peaks

during one half cycle of the input and the negative peaks during the other half cycle. The result is an accurate duplication of the modulating signal at the output. C4 filters out any local carrier components that reach the output.

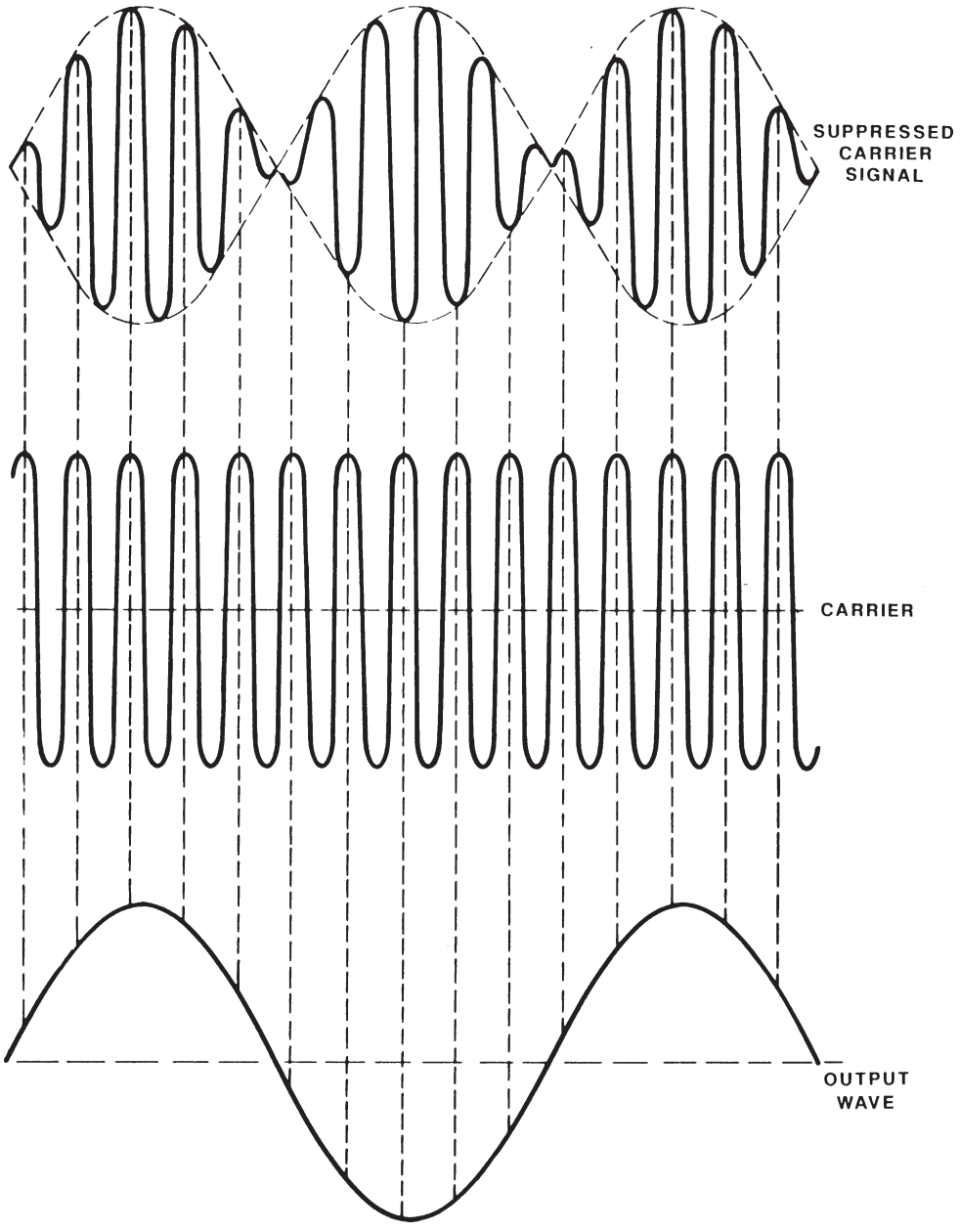


Figure 3-6

Waveforms for the synchronous detector.

You can see from the waveforms of Figure 9-6 that even slight carrier phase or frequency variation between the original carrier and the local carrier will shift the sampling point and cause severe distortion. For this reason, the local carrier signal must be held exactly in phase with the original carrier. This presents the most serious drawback to suppressed-carrier AM. Elaborate circuits are required at the receiver to generate the carrier signal in the correct phase and frequency. However, suppressed-carrier AM is used in the stereo FM and color TV broadcasting systems. In these systems, a special synchronizing signal is also transmitted. The receiver recovers this synchronizing signal and uses it to generate a local carrier signal in the correct phase and frequency.

THE PRODUCT DETECTOR

The simple diode detector is also unsuitable for demodulating a single-sideband signal. This is because an SSB waveform's envelope corresponds only to the amplitude of the modulating signal, not the frequency. In fact, if an SSB transmitter is modulated by a constant amplitude audio tone whose frequency is varied, the SSB envelope is constant but its frequency varies. Therefore, an SSB demodulator must be sensitive to both frequency and amplitude variations.

In product detection, an internal reference signal is generated and the SSB signal is compared with it. This reference signal is essentially a "substitute" carrier for the single sideband. The product detector itself is a nonlinear mixer. It can take on many forms, but is usually a balanced modulator circuit or a dual-gate MOSFET.

A typical product detector is shown in Figure 9-7. It uses a diode-ring balanced modulator as the nonlinear mixer. The internally generated carrier, which is provided by a BFO (Beat Frequency Oscillator), is applied to the center tap of T1. The incoming SSB signal is applied to the primary of T1. The two signals are then mixed across the diode ring. If the carrier signal is within approximately 50 Hz of the original suppressed carrier, the difference signal is very close to the original modulating waveform. R1, R2, C1, and C2 form an RC low-pass network which removes any carrier component and allows only the audio signal to pass to the output. In this way, the original information has been recovered.

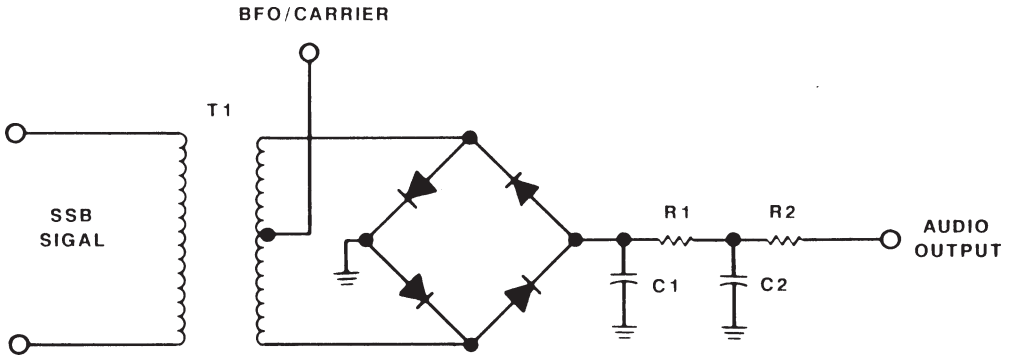


Figure 9-7
Balanced modulator product detector.

As you can see, the requirements for the internally generated carrier for SSB are not nearly as stringent as they are for double-sideband, suppressed-carrier AM. In suppressed-carrier AM, the local carrier has to be exactly the same frequency and phase as the original. In SSB, it has to be only close to the original. This greatly simplifies the SSB receiver and, along with bandwidth considerations, is the primary reason for the widespread use of SSB instead of double-sideband, suppressed-carrier AM.

Since the local carrier is not the same frequency and phase as the original, the output quality of SSB does suffer somewhat. However, the demodulated signal is still intelligible and, therefore, quite effective for voice and data communications.

SUMMARY

In this chapter you learned about the circuits that are used to detect AM signals. You learned that different techniques are required for standard AM, double-sideband suppressed-carrier AM, and SSB. You also learned that detectors look very much like modulators.

SELF-TEST REVIEW

1. The most popular AM demodulator is the _____ _____ .

2. A transistor can be used as an AM detector to provide additional receiver _____ .

3. When a transistor is used as an AM detector, it is biased for class _____ operation.

4. The AM diode detector is normally chosen over the transistor detector for its _____ and _____ .

5. The _____ detector must be used to demodulate double-sideband, suppressed-carrier AM signals.

6. What are the requirements for the local carrier signal when demodulating a double-sideband, suppressed-carrier signal?

   _____

   _____ .

7. What are the primary applications of double-sideband, suppressed-carrier AM? _____

   _____

   _____ .

8. What type of demodulator is used for single-sideband signals?

   _____

   _____ .

9. What are the requirements for the local carrier signal when demodulating a single-sideband signal? _____

   _____

   _____ .

CHAPTER 10 — Tuned RF and Superheterodyne Receivers

INTRODUCTION

When the number of transmitters increased, tuned circuits were added to receivers to provide selectivity and to reduce interference. Inevitably, it became necessary to receive signals from farther away, which were too weak to provide a useful signal in a headset. So, tuned RF amplifiers were added to improve the sensitivity and to further improve selectivity.

However, tuned RF receivers had several problems that limited their usefulness. In 1918, the invention of the superheterodyne receiver overcame many of these problems, and performed so well that the superheterodyne technique remains the fundamental form of receiver architecture today, in all forms of communications and radar equipment.

Receiver architectures are only coincidentally related to AM; in fact, all of them are useful for receiving all types of modulated signals. It just happens that AM was the predominant form of modulation in use when the technology of receiver design was developing rapidly, so receiver technology "grew up" with AM signals. As we haven't studied receivers before this point in the course, it makes sense to merge the study of receiver architecture with the study of AM detectors in this chapter, so that you can see how a system of transmitter and receiver is implemented.

OBJECTIVES

- Identify the block diagrams of a tuned RF receiver and a superheterodyne receiver.
- List the disadvantages of a tuned RF receiver.
- List the stages of a typical superheterodyne receiver and state the function of each.
- State the factors that must be considered when you select an intermediate frequency.
- Find the image frequency when you are given the received and intermediate frequencies.
- State the purpose of an AGC circuit.

TUNED RF RECEIVER

If a simple diode detector were connected directly to an antenna, there would be no means of selecting a specific signal. All signals present at the antenna would be demodulated and the output would be a jumbled mess. One way to select the desired signal is to connect a tuned RF amplifier between the antenna and detector. However, the bandwidth of a single RF amplifier is not narrow enough to reject all the unwanted signals. Therefore, several RF amplifiers are used as shown in Figure 10-1. This reduces the bandwidth to a point where only one signal passes to the detector. The RF amplifiers also increase the amplitude of the desired signal before it is applied to the detector. Thus, the additional RF amplifiers increase both the sensitivity and selectivity of the receiver.

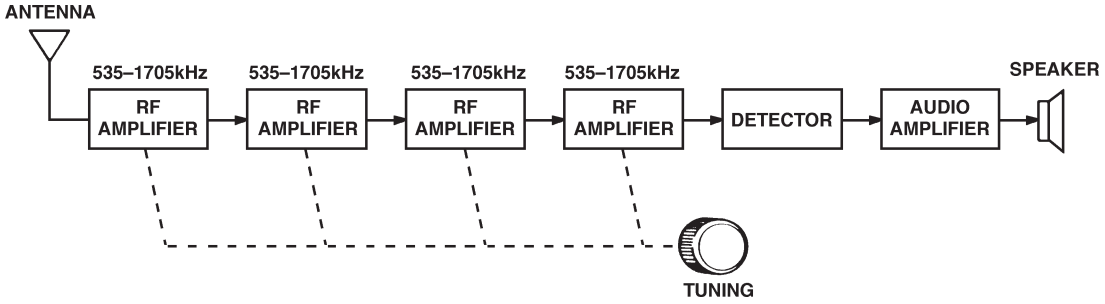


Figure 10-1

The tuned RF receiver.

Sensitivity

Sensitivity is a measurement of a receiver's ability to receive weak signals. For example, if a receiver can "pick up" an extremely weak signal, it is said to have good sensitivity. The amount of gain a receiver has determines its sensitivity. Therefore, the more gain obtained in the RF section, the greater is the receiver's sensitivity.

Selectivity

Selectivity is a measure of a receiver's ability to select one signal while rejecting all others. In a tuned RF receiver, this is determined by the bandwidth of the RF amplifiers. Therefore, as the bandwidth is reduced, selectivity is increased.

The receiver shown in Figure 10-1 has four RF amplifier stages. Its sensitivity and selectivity are adequate for the AM broadcast band of 535 to 1705 kHz. However, it is extremely difficult to get all the RF amplifiers to tune to exactly the same frequency as the receiver is tuned across the band. Another problem is that the bandwidth of the RF amplifiers increases as the operating frequency increases. As a result, the selectivity at 1700 kHz is not as good as it is at 540 kHz. It is also extremely difficult to design RF amplifiers that provide high gain and are tunable over the required range. Because of these and other problems, the tuned RF receiver was abandoned in favor of the superheterodyne receiver. Superheterodyne receivers are discussed in detail next.

SUPERHETERODYNE RECEIVERS

Most of the problems of the tuned RF receiver are associated with tuning the amplifiers across a frequency band. Operating at a fixed frequency, the tuned RF set can be made very sensitive and selective. The superheterodyne receiver capitalizes on this fact by converting all received signals, regardless of frequency, to a predetermined fixed frequency, which is then applied to a fixed-tuned RF receiver. The required sensitivity and selectivity are then obtained relatively simply (as compared to the tuned RF set) in the fixed-tuned stages.

Block Diagram

The block diagram of a typical, standard AM broadcast superheterodyne receiver is shown in Figure 10-2. The RF amplifier is similar to that used in the tuned RF receiver. It provides signal gain and a limited degree of selection. However, the majority of the receiver's selectivity and gain are obtained from the later stages. In fact, in some receivers, the RF amplifier is eliminated altogether.

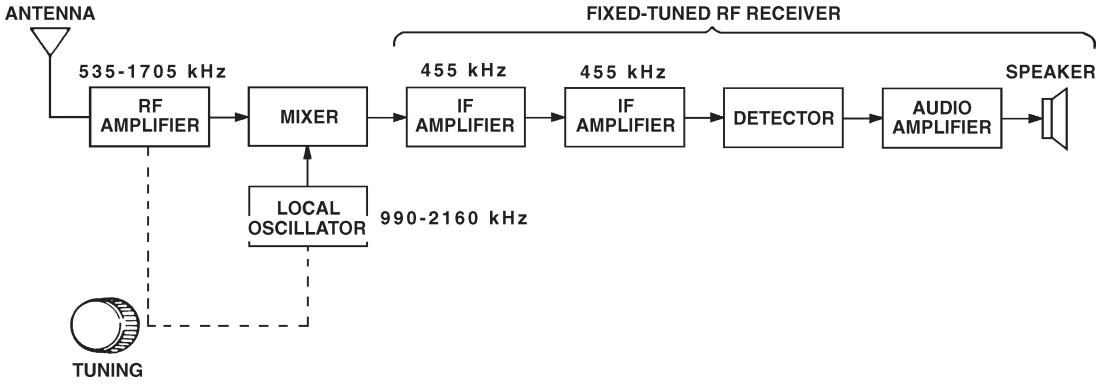


Figure 10-2

Standard AM broadcast superheterodyne receiver.

The mixer and local oscillator are the keys to the superheterodyne action. They convert the selected RF signal to a fixed frequency for further amplification. In this case, the fixed frequency is 455 kHz which is a commonly used intermediate frequency or IF. It is called an "intermediate" frequency because it is between the original radio frequency and the detected radio frequency.

To convert the incoming RF signal to 455 kHz it must be mixed, or heterodyned, with a signal that is exactly 455 kHz above or below it. That is, the difference between the selected RF signal and the local oscillator must be 455 kHz. In the receiver of Figure 10-2 the local oscillator is tuned above the incoming signal. Therefore, if the desired signal is a 1000 kHz, the tuning knob would be adjusted to 1000 kHz. This would adjust the RF amplifier to 1000 kHz, and allow a relatively small band of frequencies to be amplified and applied to the mixer. The tuning knob would also adjust the local oscillator to exactly 455 kHz above the incoming RF signal or 1000 kHz + 455 kHz = 1455 kHz.

Both the RF and local oscillator signals are then applied to the mixer. The mixer is any nonlinear circuit, and could even be a balanced modulator. The outputs of the mixer are the original signals and the sum and difference mixing products. Since the mixer's output and several following amplifiers are fixed-tuned to 455 kHz, only the difference signal, in this example 1455 kHz – 1000 kHz or 455 kHz, leaves the mixer and is amplified.

The IF amplifiers are actually fixed-tuned RF amplifiers. There are usually tuned resonant circuits on both the input and output of each stage. This assures both high gain and narrow bandwidth which means better receiver sensitivity and selectivity. Since they are fixed-tuned, their gain and bandwidth are constant. There are also no tuning problems since each stage is individually "peaked" only once, during its initial alignment at the factory. The selection of the intermediate frequency itself involves several factors which are discussed later in this chapter.

The amplified IF signal is then coupled to the detector where the original modulating information is recovered. In the standard AM receiver of Figure 10-2, the detector is a simple diode demodulator. A single-sideband receiver requires a product detector and a BFO. The detected audio is then amplified and applied to the speaker.

Frequency Conversion Circuits

A typical receiver front-end circuit is shown in Figure 10-3. Q1 is the RF amplifier, Q2 is the mixer, and Q3 is the local oscillator. The RF amplifier is tuned to the desired input frequency by C1. As shown in the schematic, C1 is mechanically coupled to C2. Therefore, C2 simultaneously tunes the local oscillator to its correct frequency which is the incoming frequency + the IF. C2 resonates with L1 to form the tank for the oscillator while L2 provides regenerative feedback.

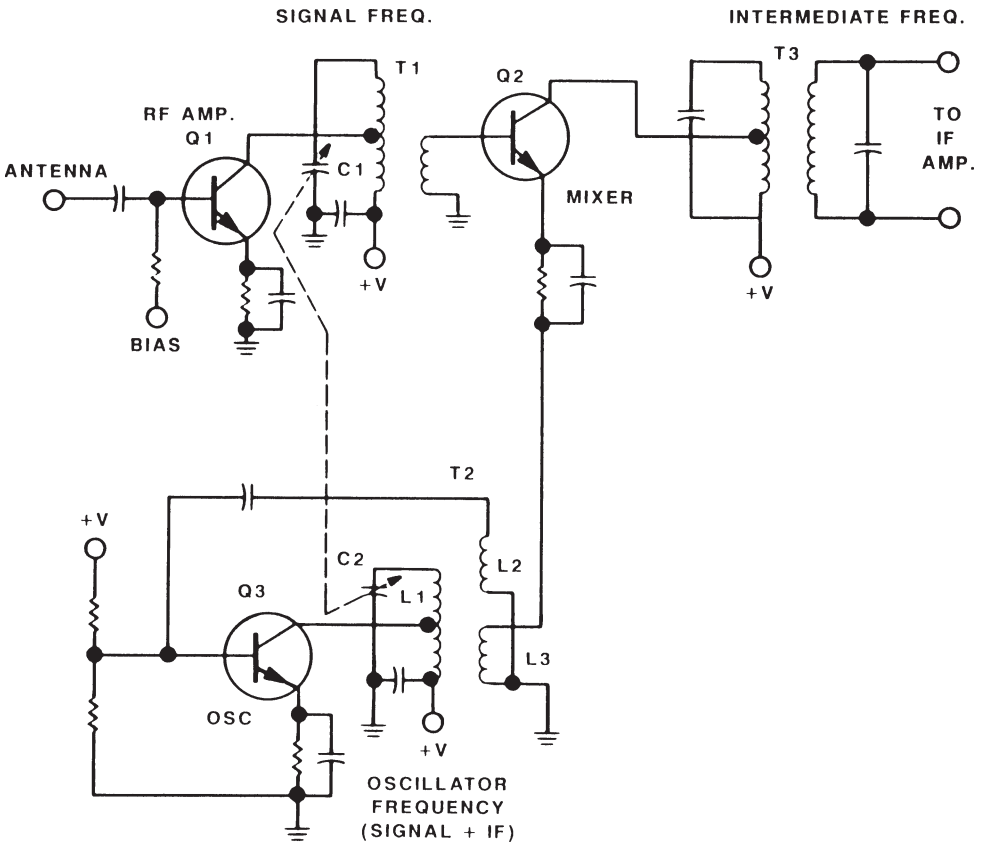


Figure 10-3

Frequency conversion circuit.

L3 couples the local oscillator signal to the emitter of Q2, while the input RF signal is coupled to the base. These two signals are then mixed across the nonlinear resistance of Q2. The output of the mixer is applied to the primary of T3 which is tuned to the intermediate frequency. Thus, the difference signal, which is the intermediate frequency, is separated and applied to the IF stages where it is further amplified.

A variation of the frequency conversion circuit is shown in Figure 10-4. In this circuit, transistor Q2 performs the dual function of local oscillator and mixer. The key to this circuit is transformer T2. Here, L1 couples the collector signal of Q2 to the resonant tank of L2 and C2. C3 then couples the oscillations of L2 and C2 to the emitter of Q2. The transistor amplifies this signal and the continued feedback of L1 sustains the oscillations. Thus, Q2 functions as an Armstrong oscillator.

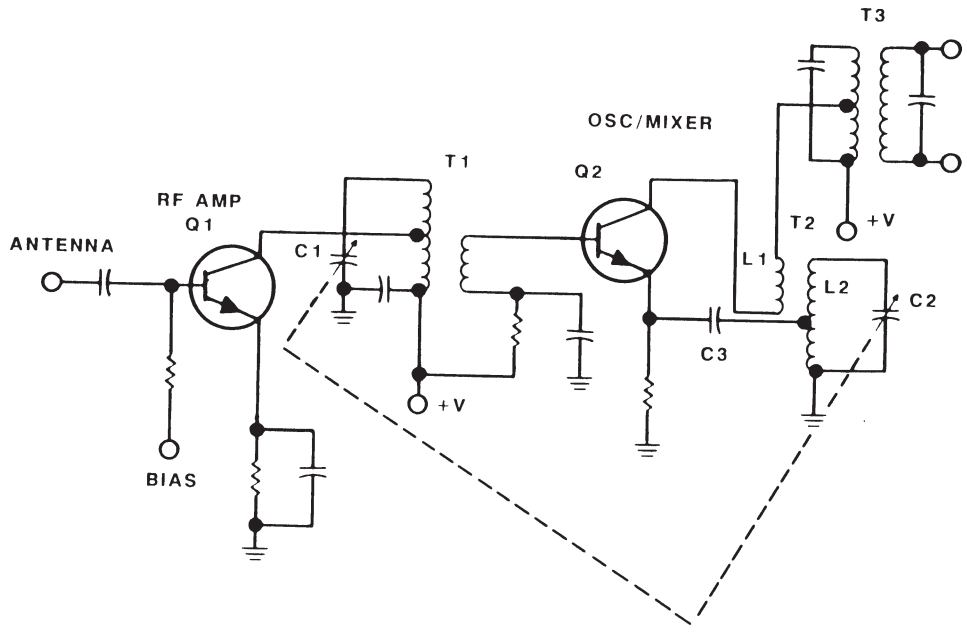


Figure 10-4

Autodyne conversion circuit.

When a signal is present at the antenna, Q1 and the tuned circuit of C1 and T1 select and amplify the desired signal. It is then coupled to the base of Q2 and mixed with the local oscillator signal. The output of Q2 is coupled through L1 to the primary of T3 which is tuned to the intermediate frequency. Thus, the difference signal is separated and applied to the IF stages. This type of circuit is called an autodyne. It is usually used in low cost portable receivers, and is often used in conjunction with the transistor detector discussed earlier. Both circuits perform dual functions and thus allow simplicity and lower cost.

A high performance mixer circuit often found in HF, VHF, and UHF receivers is shown in Figure 10-5. It uses a dual-gate MOSFET as the active element. The MOSFET has a high input impedance and, therefore, causes less loading on the previous circuit. In this mixer the RF signal is applied to gate 2 while the local oscillator signal is applied to gate 1. This minimizes any interaction between the two signals. The mixing action is produced by biasing the MOSFET into its nonlinear region. The difference output signal is then separated by the tuned primary of T2.

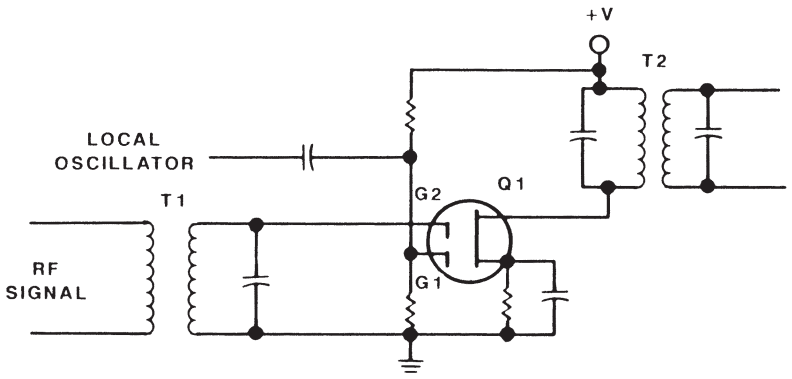


Figure 10-5
A MOSFET mixer.

Intermediate Frequency Selection

Several factors must be considered when selecting an intermediate frequency. Many of these factors revolve around the equation BW = F_o /Q. This states that bandwidth is equal to the operating frequency divided by the circuit Q. Therefore, for good selectivity, or a narrow bandwidth, the circuit Q must be high and the operating frequency low. Thus, it is generally desirable to use a relatively low intermediate frequency. Other advantages of a low IF are better circuit stability and higher gain stages. For example, as frequency increases, circuit resistive, radiation, and dielectric losses increase; stray capacitance and inductance effects increase; and stray feedback causes instability. Therefore, low intermediate frequencies are desirable.

Another factor to be considered when selecting an intermediate frequency is the required bandwidth. For example, if an AM broadcast receiver uses a 60 kHz IF with a Q of 60, the bandwidth is 1 kHz. This is hardly practical for receiving an AM signal whose bandwidth is 1 kHz, let alone an AM signal whose bandwidth is at least 10 kHz! Therefore, the system bandwidth must be taken into account.

The final factor to be considered is image frequency rejection. The image frequency is an undesired RF signal which, when mixed with the local oscillator, will also produce the correct IF value. As an example, suppose the IF is 200 kHz and the desired signal is 4.2 MHz. If the local oscillator is tuned above the incoming signal, it must be at 4.2 MHz + 0.2 MHz = 4.4 MHz. Therefore, when the local oscillator and RF signal are mixed, the difference signal will be 200 kHz, which is the IF. However, there is another signal that produces a 200 kHz IF signal. It is located 200 kHz above the local oscillator at 4.6 MHz. When this 4.6 MHz signal is mixed with the 4.4 MHz local oscillator the difference is also 200 kHz. If applied to the mixer, this image frequency is amplified just as if it were the desired signal.

Let's try another example of finding the image frequency. If the IF is 455 kHz and the received signal is 1110 kHz with the local oscillator tuned above the RF signal, what is the image frequency? The local oscillator frequency is 1110 kHz + 455 kHz = 1565 kHz. Therefore, the image frequency is 1565 kHz + 455 kHz = 2020 kHz.

From the previous examples, you can see that the image frequency is separated from the desired signal by twice the intermediate frequency. Rejection of the image signal depends solely on the selectivity of the RF amplifier and mixer stages. The higher the IF is, the farther the image signal will be from the desired signal. To improve image rejection, a higher intermediate frequency should be used. This conflicts with the previous factors considered. Therefore, the selection of an intermediate frequency must be a compromise.

Once all of these factors were taken into account, several "standard" intermediate frequencies were chosen. As an example of this, AM broadcast receivers (535 – 1705 kHz) generally use an IF of 455 kHz. This frequency provides excellent Q for high gain and good selectivity. It is also high enough to place the image frequency well outside the RF amplifier's passband. FM broadcast receivers use an IF of 10.7 MHz. This was chosen to provide sufficient bandwidth, up to 200 kHz, for the FM signal. It also places the image frequency 21.4 MHz away from the desired RF signal. Television receivers use an IF of approximately 40 MHz. This allows the extremely wide TV bandwidth of 4 MHz to be accommodated, while still leaving the image frequency outside the RF passband.

IF Amplifiers

An IF amplifier is simply a fixed-tuned RF amplifier. Figure 10-6 shows a typical amplifier tuned to 455 kHz. The only difference between this circuit and a typical RF amplifier is the double-tuned transformers T1 and T2. The resonant circuits in both the primary and the secondary improve the selectivity and circuit gain. Apart from this, the circuit is a straightforward RF amplifier with C1 providing neutralization for added stability.

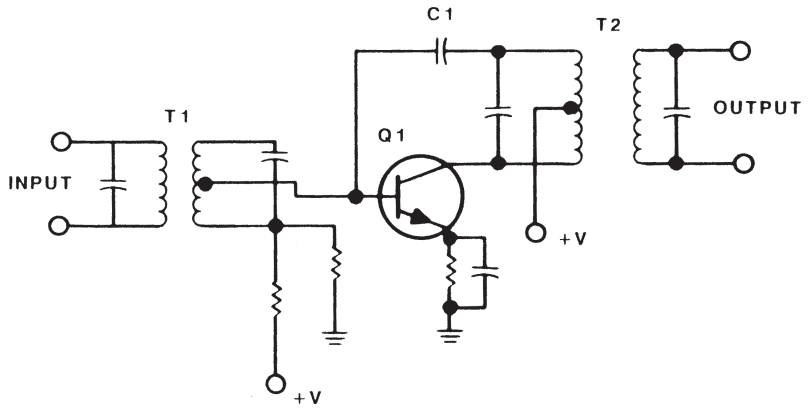


Figure 10-6
An IF amplifier.

The selectivity of a receiver can be tailored to the exact requirements by using a bandpass filter as part of the IF section. This technique is used in communications receivers where narrow bandwidths are required, in FM receivers where the high 10.7 MHz IF does not provide adequate selectivity, and in TV receivers where it eliminates tedious tuning and provides optimum IF response, and in other applications where a tailored frequency response is needed.

Automatic Gain Control

When you tune across the AM broadcast band, it is not uncommon to hear an extremely weak station at one point and a strong local station at the next frequency. To compensate for these level variations, Edwin Armstrong invented automatic volume or gain control. This circuit automatically varies the gain of the receiver to keep the volume at a relatively constant level regardless of the input signal.

A simple automatic gain control (AGC) circuit is shown in Figure 10-7. Actually it is an AM diode detector with R2 and C2 added. These components take the detected and filtered audio signal and filter it still further to remove the audio signal. What remains is a DC voltage that is directly related to the input signal amplitude. If a strong input signal is present, the AGC voltage is high. Just the opposite is true for a weak signal. This voltage is then used to control the receiver gain. While a positive AGC voltage is shown in Figure 10-7, negative voltage is also used and is obtained by reversing the polarity of D1.

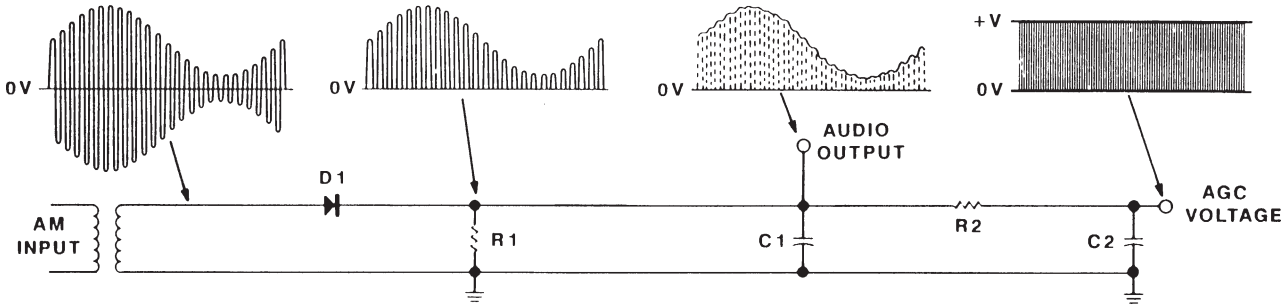


Figure 10-7

A simple automatic gain control circuit.

To see how the AGC voltage controls receiver gain, consider the block diagram shown in Figure 10-8. Here, the DC voltage derived from the detector is applied to both IF amplifiers and the RF amplifier. This voltage is used to bias these stages. By varying the stage bias, the gain of each stage can be controlled. For example, when a strong signal is present, the AGC voltage is high and it decreases the gain of the RF and IF amplifiers. The result is a relatively constant audio output level. You can see that the AGC circuit is a valuable addition to a receiver. However, some modes of operation such as SSB do not have constant signals as with AM. This makes it difficult to derive an AGC voltage. In this case, special AGC circuits have been developed.

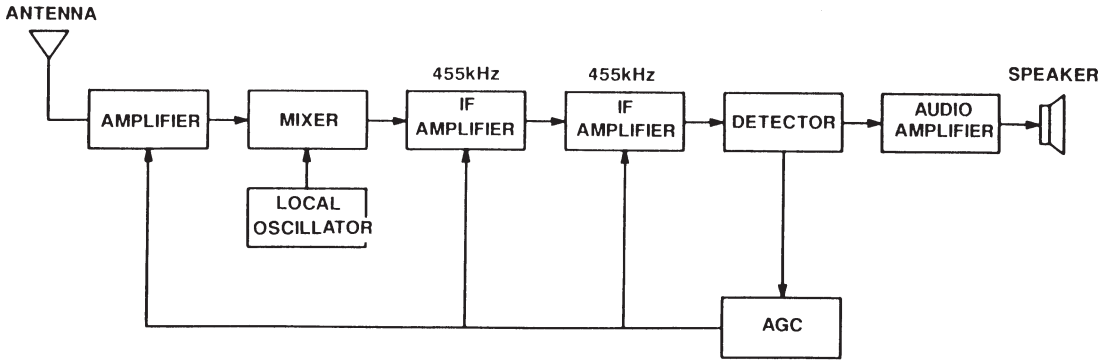


Figure 10-8

A superheterodyne receiver with AGC.

An AGC circuit used with SSB receivers is shown in Figure 10-9. Since there is no carrier present in SSB, the AGC system must derive its voltage from some other part of the signal. The circuit of Figure 10-9 uses the demodulated audio to obtain the AGC voltage. Here, the audio is amplified by the operational amplifier and then rectified by the voltage doubler circuit of D1/D2 and C1/C2. Therefore, the DC voltage across C2 is directly proportional to the strength of the SSB signal. However, if no audio is transmitted this voltage drops to zero. For example, in between words and sentences, when the speaker pauses, C2 attempts to discharge through R1. The time constant of C2/R1 is chosen to prevent the AGC voltage from dropping during normal pauses in speech.

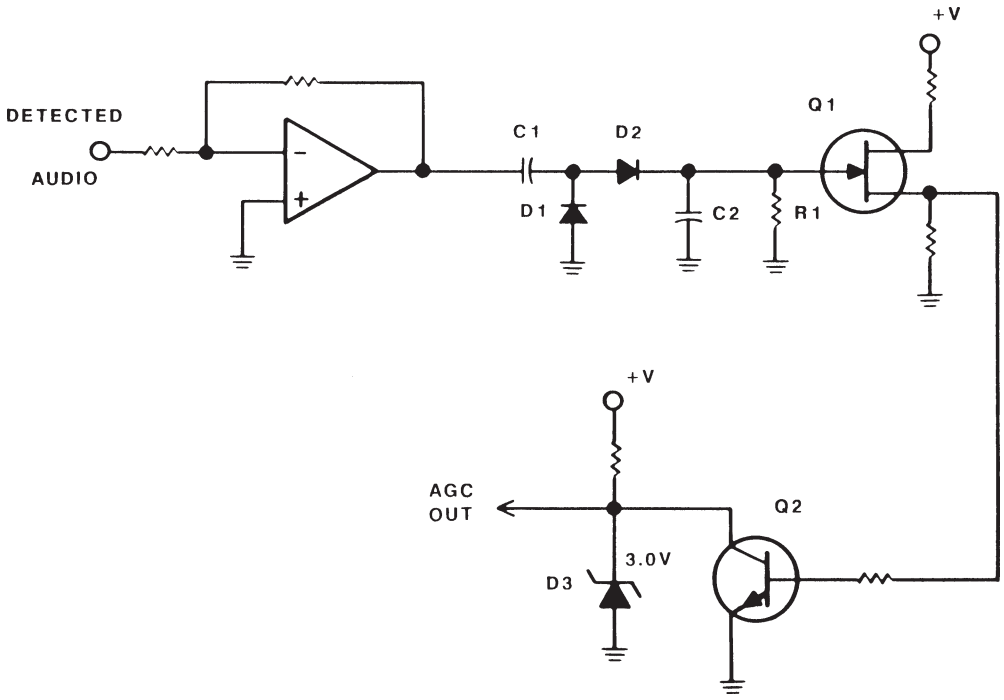


Figure 10-9

Audio derived AGC circuit.

JFET Q1 and transistor Q2 form an interface between the AGC detector and the receiver circuits being controlled. Q1's high input impedance allows it to amplify the DC voltage across C2/R1 without affecting the time constant. The output of Q1 is applied to the base of Q2 and when the voltage is sufficient, it turns on Q2. Up until this time the AGC voltage is held at a constant 3 V by zener diode D3. When Q2 turns on, it effectively turns off the zener by reducing the voltage below 3 V. Q2 now controls the AGC voltage and any increase in the detected signal at C2 causes Q2 to conduct more, which lowers its collector voltage, and, hence, the AGC voltage. Thus, as the received signal increases, the AGC voltage decreases, which reduces the receiver gain and holds the audio output relatively constant.

Double Conversion Receivers

As mentioned earlier, choosing an intermediate frequency involves a compromise. A low IF is desired because of its better gain, stability, and narrow bandwidth. However, a high IF is desirable to improve image rejection. The double-conversion receiver is designed to take advantage of both high and low IFs by using a high "first" IF and then converting to a low "second" IF.

A simple double-conversion receiver is shown in Figure 10-10. This receiver uses two mixers, two local oscillators, and two intermediate frequencies. The first IF is selected to place the image as far away from the desired signal as possible. In this case, the 9 MHz first IF places the image 18 MHz away from the desired signal. Therefore, the image is well outside the bandwidth of the RF amplifier and is greatly attenuated.

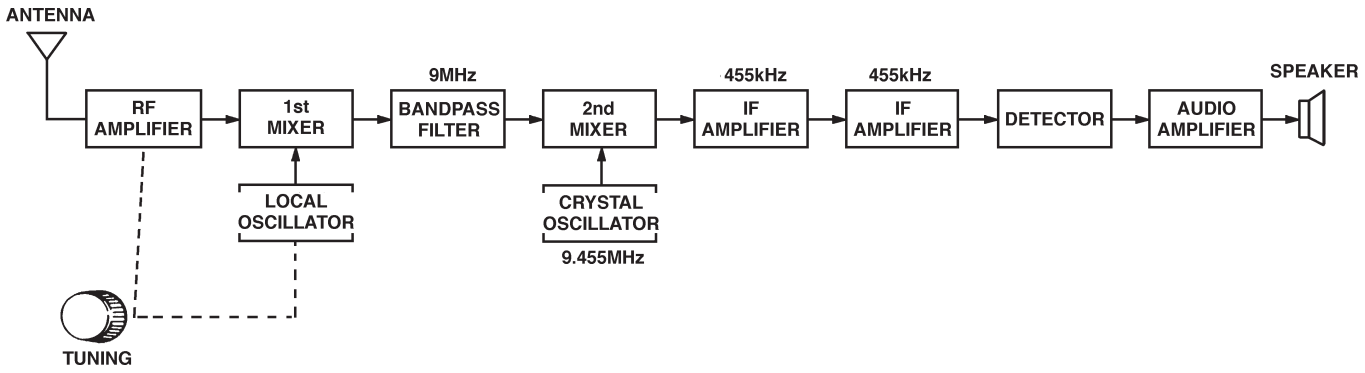


Figure 10-10

Double conversion receiver.

To demonstrate how the receiver operates, we must follow the signal from the antenna to the speaker. As an example, suppose you want to receive the 5 MHz signal of the standard time and frequency station, WWV. The tuning knob is set to 5 MHz, which tunes the RF amplifier to 5 MHz and the local oscillator to 14 MHz. This places the image frequency at 23 MHz, well outside the bandpass of the RF amplifier.

The incoming signal is amplified and then applied to the first mixer along with the local oscillator signal. At the first mixer output, the difference signal of 9 MHz is selected by the bandpass filter and applied to the second mixer. The bandpass filter may be an LC filter or a crystal filter for improved selectivity. In this receiver, there is no first IF amplifier since most of the receiver's gain is supplied by the second IF.

Since a 9 MHz signal always appears at the second mixer, a crystal controlled "second" local oscillator is used. Its frequency of 9.455 MHz is mixed with the 9 MHz input signal and the difference frequency, 455 kHz, is selected by the IF amplifiers.

The IF amplifiers increase the signal amplitude and apply it to the detector. The detector demodulates the signal and recovers the original audio. This audio signal is then amplified and applied to the loudspeaker.

Other Applications

The superheterodyne receiver, although invented in 1918, remains the basic and most versatile form of communications receiver. As a result of its superior performance, the superheterodyne is used in virtually every kind of electronic communications system.

The ability to convert received signals to a standard frequency for processing makes this a versatile architecture. As an example, it is difficult to obtain high gain and selectivity at the UHF and microwave frequencies which are used for many point-to-point communications systems. However, using superheterodyne technology, the high signal frequency can be quickly converted to a lower intermediate frequency where it can be readily processed as required. Frequently, the bandpass characteristics of the IF amplifiers are specifically tailored to the needs of the application, providing communications capabilities that would not be possible without superheterodynes.

To deal with the opposing requirements posed by images and sensitivity and selectivity, multiple-conversion superheterodynes have also been constructed, which use three, and occasionally four, mixers

before reaching the final IF. Superheterodyne architecture is commonly used in receivers that terminate optical communications links, where the carrier frequency is so high the wavelength is measured in fractions of a millionth of a meter.

It is common to find superheterodynes in fixed communications systems that, although tunable, are normally tuned to a specific frequency when installed and left at that frequency. Even though a tuned RF receiver would theoretically be usable in such a situation, superheterodynes are so much superior that it is rare to find another receiver type used.

Nothing about superheterodynes makes them suitable only for AM; the kind of detector chosen determines the type of signals the receiver will process. Until the detector is reached, all signals look alike to the receiver, and all are treated the same way. So, you will find this same superheterodyne architecture in pocket AM radios, radar receivers, and telephone system satellite terminals, each of which uses a different type of modulation. In later chapters, we will describe how to detect other types of modulation. Those detectors can be "plugged into" the superheterodyne architecture, and everything will work perfectly.

It is a common and expected practice to tailor the receiver design for the type of signals it is expected to handle. Thus, you will find differences in design details of receivers used for AM broadcast, FM communications, data, radar, etc. In the next chapter, we will look in detail at how receiver performance is specified, and what imperfections must be taken into consideration when receivers are selected for a given application.

SUMMARY

In this chapter, you learned about the architecture of receivers, without which there would be no signals to detect. You examined the tuned RF receiver and discovered that it presents many practical problems for the designer. You then studied the superheterodyne architecture in detail, and found that it could resolve most of the problems associated with the tuned RF system.

Finally, you learned that receiver architectures are not bound to the type of signal to be received. Whether AM, FM, PM, or something else, the receiver architecture you find is almost always the superheterodyne, with only the detector changed to correspond with the type of modulation being used.

SELF-TEST REVIEW

1. A tuned RF receiver uses one or more _____ in front of the detector.

2. What is receiver sensitivity? _____

   _____

3. What is receiver selectivity? _____

   _____

4. List the disadvantages of the tuned RF receiver. _____

   _____

   _____

5. Write in the names of the blocks of the superheterodyne receiver shown in Figure 10-11.

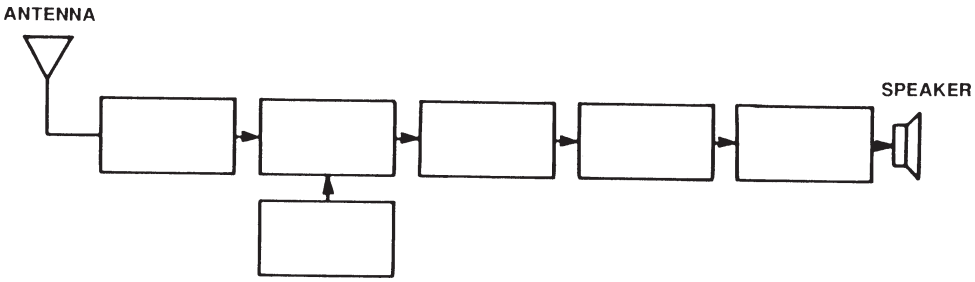


Figure 10-11

Figure for Questions 5 and 6.

6. State the function of each stage in Figure 10-11. _____

   _____

   _____

7. What is an autodyne circuit? _____

   _____

   _____

8. Dual-gate MOSFETs are often used in superheterodyne receivers as _____ .

9. What factors must be considered when selecting an intermediate frequency? _____

   _____

10. What is the image frequency when the receiver is tuned to 9.77 MHz, the local oscillator is tuned above the desired signal, and the IF is 455 kHz? _____ MHz.

11. What are the standard IFs for the following types of receivers?

    AM broadcast _____

    FM broadcast _____

    TV broadcast _____

12. An IF amplifier is actually a _____ .

13. What is the purpose of a bandpass filter in the IF stage of a receiver? _____

    _____

14. What is the purpose of an AGC circuit? _____

    _____

15. Why is an AGC circuit for SSB more complex than that for standard AM? _____

    _____

16. What is the purpose of using two IFs as in a double-conversion receiver? _____

    _____

CHAPTER 11 — Receiver Analysis

INTRODUCTION

In this chapter you will learn about several common specifications that directly affect receiver performance. With these in mind, you can more easily compare one receiver against another, and pick the one that best suits your needs.

Next we will take a look at a typical communications receiver and follow the signal path through its circuitry.

OBJECTIVES

When you complete this chapter, you will be able to:

- Define receiver sensitivity and selectivity.
- Define at least six receiver specifications.
- Analyze the operation of a complex receiver from its schematic diagram and draw a block diagram of the receiver.

RECEIVER SPECIFICATIONS

The importance of specifications depends upon the application. For broadcast listeners in the 1920's and 30's the most important receiver specification was sensitivity. The receiver had to be sensitive enough to "pick up" the weak signals of the relatively few transmitters that were "on the air." However, in the 1940's and 50's there were many more transmitters, and receiver selectivity became much more important. Today, there are literally millions of transmitters and many new communications applications. While sensitivity and selectivity are still important, many new specifications are recognized as important receiver performance measurements. Let's look at these specifications in detail.

Sensitivity

Sensitivity is a measure of the amplitude of the weakest signal that can be received, with a specified signal-to-noise ratio. That is, the signal must be stronger than the noise level by a certain ratio.

Receiver sensitivity is actually a measure of how much signal it takes to overcome the internal receiver noise. This is the reason for using a specified signal-to-noise ratio. In fact, signal-to-noise ratio measurements have been replaced by a more correct measurement of signal-plus-noise-to-noise ratio. This ratio takes into account that the noise is still present, along with the signal. The signal is added to the receiver noise, and must overcome it.

Receiver sensitivity is determined by first measuring the noise output with no signal present. The output from a signal generator is then applied to the receiver input. The generator's output level is increased until the receiver's output increases by 10 dB. Once this is achieved, the signal generator's output voltage is measured. Typical radio receiver values are 1.0 to 0.5 μV of signal for a +10 dB signal-plus noise-to-noise ratio [10dB(S+N)/N]. A μV value means the receiver requires approximately 1.0 μV of signal to overcome the internal noise by 10 dB and to provide adequate reception.

You can see from the equation in the preceding paragraph that there are only two ways to improve sensitivity: increase the signal strength, or reduce the noise. Both approaches have been used.

Signal strength can be improved by reducing the loss between the transmitter and receiver. This can mean a higher power transmitter, higher gain antennas at the transmitter and receiver (if a radio system), or lower loss transmission media such as cable or waveguide.

Noise can only be reduced by careful attention to receiver design. Some designs and components are noisier than others, and must be avoided when low-noise receivers are required. Low temperature also relates to low noise. It is common to find receivers that must deal with very weak signals, such as satellite receivers, where the receiver RF amplifier and mixer are refrigerated. By cooling these critical components to very low temperatures, it is possible to amplify the signal significantly before too much noise is added to it. And there are costs associated with all of these approaches. As a result, receivers are designed to offer enough sensitivity to do their assigned job, but not to provide many times more than is needed.

There is another reason for avoiding excessive sensitivity overload. Excessively sensitive receivers have a difficult time dealing with strong signals, even if those aren't the ones the receiver is intended to pick up. It is very difficult to keep very strong signals which are close in frequency to the desired signal from getting into the "front end" (the RF amplifier and mixer). When a receiver front end is overdriven by strong signals, it is referred to as being overloaded. The amplifiers and mixers are driven to clipping, and distortion results which causes great problems in retrieving the desired signal from the resultant mess. In this regard, sensitivity is like ice cream; you can have too much of a good thing.

Selectivity

Another important receiver specification is selectivity. This measures the receiver's ability to select a desired signal and reject all others. The only way to measure selectivity is to determine the receiver's bandwidth. A typical receiver selectivity curve is shown in Figure 11-1. This shows that the receiver bandwidth is 3 kHz at the –6 dB point. That is, at 1.5 kHz either side of the center of the passband, the received signal's amplitude drops 6 dB. At 3 kHz either side of the center, the signal drops 60 dB. Therefore, the bandwidth at –6 dB is 3 kHz, and at –60 dB, is 6 kHz. The selectivity curve is primarily due to the design of the IF section.

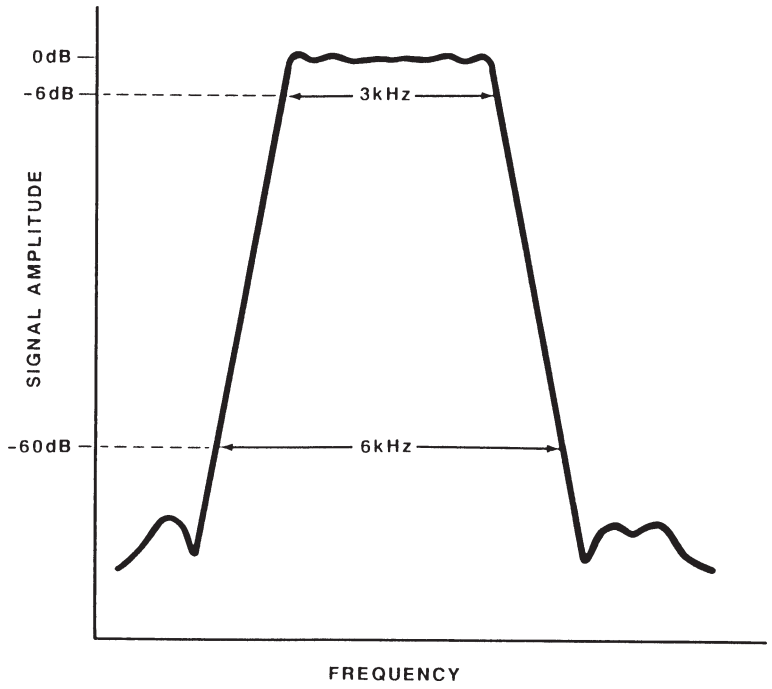


Figure 11-1

Typical selectivity curve.

Selectivity is another parameter that has to be "just right" for the application. If the bandwidth is too narrow, all of the desired signal does not pass to the detector, and information is lost. On the other hand, if the bandwidth is too wide, many signals pass to the detector, which makes it hard to retrieve the desired signal from among the multitude. In all likelihood, information will be lost in that situation, too.

There is no magical "right" number for bandwidth. To adequately detect an AM broadcast signal, the IF bandwidth must be in the vicinity of 10 kHz. For an FM signal, however, the required width is in the vicinity of 200 kHz. For television, it is nearly 5 MHz. Telephone systems using satellites often have bandwidth requirements of several tens of megahertz. But only a few hertz of bandwidth are needed to pass a low speed data signal like Morse Code. So you see, it all depends upon the application.

It is tempting to try to design receivers with selectivity curves that look like rectangles, with the same bandwidth at –60 dB as they have at –6 dB. The skirts of the selectivity curve can be made very steep, but not vertical. However, there is a price beyond the cost of the filter to pay for that steepness. As the selectivity curve's skirts get more and more vertical, the amount of phase distortion in the passband increases dramatically. This distortion can be a significant problem for some kinds of communications signals, such as digital data. As the steepness of the selectivity curve increases, so does the monetary cost of achieving that level of selectivity. Clearly, it is seldom beneficial to pay for selectivity you don't need.

Once again, receivers are normally designed with enough selectivity to do their intended job, but not more than is required. The width and steepness of the selectivity curve very much depends upon the application where the receiver is used, and are important factors in the design of the apparatus.

Image Rejection

Image rejection measures the receiver's ability to reject a strong signal at the image frequency. To measure image rejection, connect a signal generator to the receiver and set it to the image frequency. For example, if the receiver's first IF is 9 MHz and you tune the receiver to a frequency of 5 MHz, this places the image at 5 MHz + 18 MHz or 23 MHz. So you would set the signal generator to 23 MHz. Then you continue as though you were making a receiver sensitivity measurement. You increase the generator's output level until you achieve a 10 dB (S + N)/N ratio. Comparing the signal generator's present output level to the normal sensitivity of the receiver gives the image rejection.

As an example, suppose the receiver's normal sensitivity (E_1) is 1 V for a 10 dB (S + N)/N ratio and it takes 100 V at the image frequency (E_2) for a 10 dB (S + N)/N ratio. The image rejection is:

$$\text{Image Rejection} = 20 \text{ Log } \frac{E_1}{E_2}$$

$$= 20 \text{ Log } \frac{100 \text{ V}}{1 \text{ V}}$$

$$= 20 \text{ Log } 100$$

$$= 20 \times 2$$

$$= 40 \text{ dB}$$

In this example, the image rejection is 40 dB, which means the received and detected image signal will be 40 dB below the desired signal. Image rejection is basically a measure of the selectivity of the RF amplifier.

Image rejection is an important parameter primarily for receivers which are used in over-the-air systems, because you don't have much control over the signals that are present at the receiver antenna in such systems. On the other hand, in systems that are connected by closed means, such as cable or waveguide, image rejection is less important, as you usually know quite accurately what signals are presented to the receiver, and can design the system so that the image is not one of them.

IF Rejection

IF rejection is a measure of the receiver's ability to reject a strong input signal at its own intermediate frequency. For example, if the receiver's IF is 455 kHz, an input signal at 455 kHz should be rejected.

To measure IF rejection, you must set the signal generator to the receiver's intermediate frequency and increase its output level until you achieve a 10 dB (S + N)/N ratio, then compare the generator's voltage output to the receiver's normal sensitivity, just as in the image rejection measurement. Typical IF rejection ratings are 40 to 60 dB. This means that a signal at the receiver's IF will be 40 to 60 dB below the desired signal. This is an important parameter if signals at or near the IF frequency exist near the receiver.

Dynamic Range

Dynamic range is a measure of the receiver's ability to receive both very weak and extremely strong signals without introducing distortion to either. It is a function of the RF amplifier's linearity and the ability of the AGC system to control receiver gain.

To measure dynamic range, measure the sensitivity of the receiver for a 10 dB (S + N)/N ratio. This is the weak signal limit. Then increase the input signal until the output starts to distort. This marks the strong signal limit. The ratio between the weak signal limit and strong signal limit in dB is the dynamic range. Typical values range from 100 to 120 dB and greater.

An important parameter for broadcast and fixed-station receivers, dynamic range is usually less critical for wired systems. Few receivers, however, enjoy a constant received signal strength, so dynamic range must always be evaluated in light of the specific receiver application.

Noise Figure

The noise figure of a receiver tells you more than the sensitivity specification. The noise figure measures the amount of noise generated within the receiver. Noise figure is measured by dividing the receiver's input signal-to-noise ratio by its output signal-to-noise ratio, and can be expressed either in pure numbers or in decibels (which is more common). All other factors being equal, the receiver with the lower noise figure is preferred, as it contributes less noise to the output signal.

Overall Gain

The overall gain of a receiver is a measure of whether it can deliver a strong enough signal at its output for the desired application, given a specified input signal level. For example, a receiver might be specified to produce an output signal of 0.15 watts for an input signal strength of 1.2 V. This indicates the overall gain available from the receiver.

This parameter is important in selecting a receiver, as you must be sure that the set you choose can produce enough output at the expected input levels to be useful. If, in the example, you knew the received signal would be in the range of 1.2 V, but you needed a 0.5 watt output for your application, this receiver would not be suitable for that use.

Distortion

Distortion specifications are a measure of the degree to which the receiver adds distortion to the received signal as it is processed. It is usually specified in terms of percent total harmonic distortion, measured at one or more specified input signal levels. It is desirable for a receiver to add as little distortion as possible to the signal. It is also desirable that the distortion performance not change too much as the input signal is changed. If distortion rises rapidly as the input signal increases, the receiver is probably operating too close to its design margin.

AGC Performance

Many communications applications require a stable output signal over a range of input signal variations. While it is desirable that the output volume not change too much on an AM broadcast receiver between strong and weak stations, it is critical in many fixed-station applications that the output stay relatively constant if the input varies. AGC specifications identify this receiver characteristic, and are usually defined by indicating the output signal level at several input signal levels with the AGC on.

As you have seen, there is a wide variety of specifications used to describe the functioning of receivers. There is no "ideal" receiver; what is superb in one application is inadequate in another setting. Thus, specifications must be used judiciously to ensure that the properties you need are available and affordable.

ANALYSIS OF A COMMUNICATIONS RECEIVER

In this section, you will examine a communications receiver in detail. The unit chosen is the Heathkit HR-1680 HF Amateur Radio Receiver shown in Figure 11-2. It is designed to receive SSB and CW (Morse Code) transmissions. An amateur receiver must perform under the most demanding situations. This is because the amateur bands are extremely crowded with both weak, distant stations and extremely strong, local stations. Therefore, its sensitivity, selectivity, and dynamic range are constantly put to the test.

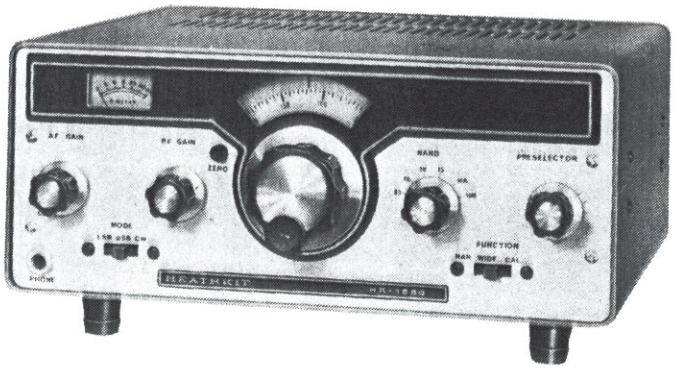


Figure 11-2
The Heathkit HR-1680 SSB/CW Receiver.

Specifications

Some of the specifications for the receiver shown in Figure 11-2 are given in Table 11-1. The first of these, frequency coverage, shows the frequency ranges that the receiver covers. It is an amateur radio receiver covering the HF "ham" bands from 3.5 MHz to 29 MHz.

Table 11-1

HR-1680 specifications.

Frequency Coverage (Megahertz)	3.5 to 4.0, 7.0 to 7.5, 14.0 to 14.5, 21.0 to 21.5, 28.0 to 28.5, 28.5 to 29.0
Sensitivity	Less than 0.5 microvolt for 10 dB signal-plus-noise to noise ratio for SSB operation.
IF Selectivity	2.1 kHz minimum at 6 dB down, 7 kHz maximum at 60 dB down.
Overall Gain	Less than 1.5 microvolt input for 0.25 watts of audio output.
Dynamic Range	120 dB or greater.
Image Rejection	50 dB or better.
IF Rejection	60 dB or better.

The sensitivity is rated at 0.5 V for a 10 dB (S + N)/N ratio, which is excellent. The IF selectivity is rated at 2.1 kHz at –6 dB and 7 kHz at –60 dB. This is adequate selectivity for SSB operation. However, it is not as good as the best mechanical filters offer, which is 2.1 kHz at –6 dB and 4.2 kHz at –60 dB. However, mechanical filters are costly items; this receiver uses a simple crystal filter to keep its cost down.

The overall gain of the receiver is an interesting specification. It shows that, while the receiver is very sensitive, it also has enough gain so that you can hear these weak signals. In this case, it states that for a 1.5 µV input, the audio output is 0.25 W.

The dynamic range specification shows that the receiver can handle an extremely wide range of signal levels without distortion. In this case, a range of up to 120 dB or greater can be handled by the receiver.

The specifications also show that image rejection is 50 dB or greater and IF rejection is 60 dB or better.

Circuit Analysis

For this discussion, you must find the envelope of schematic diagrams and remove diagram #1 for the HR-1680 Receiver.

First, let's follow the signal path which is marked by a red dashed line. The input signal first appears at the antenna and passes through the antenna terminal to a series of "traps" on the front end circuit board. These traps are tuned circuits which filter out or suppress certain undesired signals. TC101 is a series-tuned trap that shorts out any signals at the second IF of 3.395 MHz. L102 and C103 form a parallel resonant circuit that traps out signals above 30 MHz. Inductors L103, L104, and L105 form traps that attenuate any signals at the first IF of 8.395 to 8.895 MHz. The purpose of these traps is to improve both image rejection, above 30 MHz, and IF rejection.

The input signal is then coupled to Q101 (the RF amplifier) through either L106, L107, L108, or L109. L106 through L108 are tuned coupling circuits that are selected by the diode switches via the band switch. The correct circuit not only tunes the RF amplifier to the correct input frequency, but also matches the low impedance antenna to the high impedance MOSFET amplifier.

The RF signal is applied to gate 1 of the MOSFET while the dashed green line shows that AGC voltage is applied to gate 2. We'll discuss the AGC circuit in detail later. Q101's gain can also be controlled by potentiometer R1, the RF gain control. It varies the MOSFETs gain by adjusting the source voltage. The output tank circuit for the RF amplifier is chosen by diode switches via the bandswitch. These are parallel resonant tanks L111, L112, L113, L114, and L115.

The signal is then coupled to gate 1 of Q102, the first mixer. A signal from the HFO, or heterodyne frequency oscillator, is applied to gate 2. The HFO is a crystal-controlled oscillator that is used to convert the incoming signal to the first IF of 8.395 to 8.895 MHz. Actually it converts an entire 500 kHz band of signals to the first IF. However, some selectivity of the desired signal is provided by the preselector control which peaks the RF amplifier's response.

The output of the first mixer is applied to an LC bandpass filter consisting of L116, L117, and L118. This filter ensures that only the first IF of 8.395 to 8.895 MHz passes on to Q103, the second mixer. Note that the first IF consists of only the LC filter. It is used strictly to improve image rejection with most of the receiver's gain and selectivity coming from the second IF.

The signal is now coupled to gate 1 of Q103, the second mixer, while the VFO, variable frequency oscillator, signal is coupled to gate 2. At this stage, the exact received signal is selected. The first mixer and HFO selected a 500 kHz band, now the second mixer and VFO select precisely the desired signal. The difference between the VFO frequency and the desired signal's frequency is 3.395 MHz, the second IF. L119, C153, and C154 form a resonant circuit tuned to 3.395 MHz, which selects the difference signal and couples it to Q207. Q207 amplifies the signal and passes it to the crystal filter. Y203, Y204, Y201, and Y202 form a relatively simple crystal filter which provides adequate selectivity at low cost. The crystal filter's response determines the receiver's selectivity.

The output of the crystal filter is coupled to gate 1 of Q205, the IF amplifier. The dashed green line shows that AGC voltage is applied to gate 2. Therefore, the IF amplifier's gain is varied in response to the input signal level.

After the IF amplifier, the signal passes to the product detector, IC204, which is an MC1496 integrated circuit balanced modulator. In this case, the RF signal is mixed with a signal from the lower sideband BFO or the upper sideband BFO. Both of these BFOs are crystal controlled since the RF signal is always at the IF of 3.395 MHz. The output of the product detector is the difference between the BFO and the RF signal, which is the demodulated audio.

The audio is now amplified by op amp IC203C. At the output, it goes on to the function switch which selects either wide bandwidth audio or narrow bandwidth audio. If wideband audio is selected, it goes on as shown by the red dashed line. However, if narrow band audio is selected, the output of IC203C continues on to IC203A and IC203B which form an active bandpass filter. The bandwidth of the active filter is 300 Hz centered on 750 Hz. This provides additional selectivity when receiving CW signals (Morse Code). Finally, IC202 is the audio power amplifier which drives the speaker or headphones.

AGC Circuit

The AGC circuit in this receiver is very similar to that discussed earlier in this unit. Its signal path is shown by the dashed green line. The demodulated audio signal is taken just after the product detector, and amplified by IC203D. The amplified audio is then rectified by the voltage doubler formed by D205 and D206. This DC signal is applied to Q204. However, by following the solid green line, you'll see that the mode switch selects a specific resistor and, therefore, a specific time constant for the AGC system. In this case, a high value resistor, 680 kΩ, is chosen for SSB operation and a low value, 47 kΩ, for CW operation. Thus, the AGC circuit holds its control voltage longer on SSB to compensate for speech pauses. However, on CW the AGC drops out quickly to allow rapid transmit-receive Morse Code operation.

FET Q204 provides a high input impedance interface to the delicate RC time constant of the AGC system. The output on the source of Q204 passes to the base of Q203. It is also used to drive the S meter to give an indication of received signal strength. Q203 controls the AGC line. However, for low signal levels the AGC line is held at a constant 2.8 V by the zener diode. When the rectified AGC signal is high enough, Q203 turns on and takes control of the AGC line by reducing the voltage past the zener voltage. Therefore, a large input signal turns on Q203, lowering the AGC line and reducing the receiver gain. It reduces the receiver gain by lowering the gain of Q101, the RF amplifier, and Q205, the IF amplifier.

SUMMARY

In this chapter you learned about the specifications that are used to describe receiver performance, and how they apply to certain applications.

Next you followed the signal flow through a typical communications receiver from the antenna to the speaker or headphones.

Last you studied the AGC signal path and learned how switching a resistor value changes the AGC time constant so the receiver can accommodate both speech and Morse Code operation.

SELF-TEST REVIEW

1. Define the following terms as they refer to receiver specifications:

   Sensitivity: _____

   _____

   Selectivity: _____

   _____

   Image Rejection: _____

   _____

   IF Rejection: _____

   _____

   Dynamic Range: _____

   _____

   Noise Figure: _____

   _____

2. Draw a block diagram of the HR-1680 communications receiver using its schematic diagram. Be sure to include the AGC system.

Section 4

Angle Modulation

CHAPTER 12 — Angle Modulation

INTRODUCTION

In this chapter, we will build on your knowledge of modulation and transmitters by investigating angle modulation. We will first study how angle modulation is accomplished, and then we'll look at how to construct a transmitter using that knowledge. To put together all of the concepts studied, we will examine the schematic diagrams of two typical angle modulation transmitters.

Recall that angle modulation covers both frequency and phase modulation. In this chapter, you will study both forms of modulation, and we will introduce a new type of detector: the feedback demodulator. Angle modulation, especially forms of phase modulation, is widely used in modern communications systems, so it is important that you thoroughly understand the concepts presented in this chapter before you proceed.

There are two ways to achieve angle modulation. The "direct" modulation method involves directly varying the frequency of an oscillator in accordance with the modulating signal. "Indirect" modulation uses the modulating signal to vary the phase of a carrier signal that has been generated by a stable oscillator. Depending upon the application, there is a place for each of these approaches, as you shall see in this chapter.

OBJECTIVES

When you complete this chapter, you will be able to:

- Name the two types of angle modulation and three different modulator circuits.
- Draw a block diagram of an angle modulation transmitter.
- Find the master oscillator frequency and the required oscillator deviation when given the transmitter's multiplication, output frequency, and output deviation.
- Draw the block diagram of an FM receiver, transmitter, and transceiver when given their schematic diagrams.

REACTANCE MODULATOR

The first type of direct modulation we'll describe is called "reactance modulation". Here, an FET or transistor stage is biased so that it presents a capacitive or inductive reactance to the master oscillator. When a signal is applied to the modulator, its reactance varies and, therefore, so does the master oscillator's frequency.

A typical reactance modulator is shown in Figure 12-1. In this circuit, C1 and R1 are connected directly across the master oscillator's tank circuit. The reactance of C1 at the oscillator frequency is at least six times greater than the value of R1. Therefore, the C1/R1 parallel branch appears almost purely capacitive. Thus, the voltage applied to the gate of Q1 leads the oscillator voltage by 90°, due to the action of C1/R1. This in turn causes the RF component of Q1's drain current to also lead the oscillator voltage by 90°. Therefore, the entire circuit appears as a capacitive reactance across the master oscillator tank.

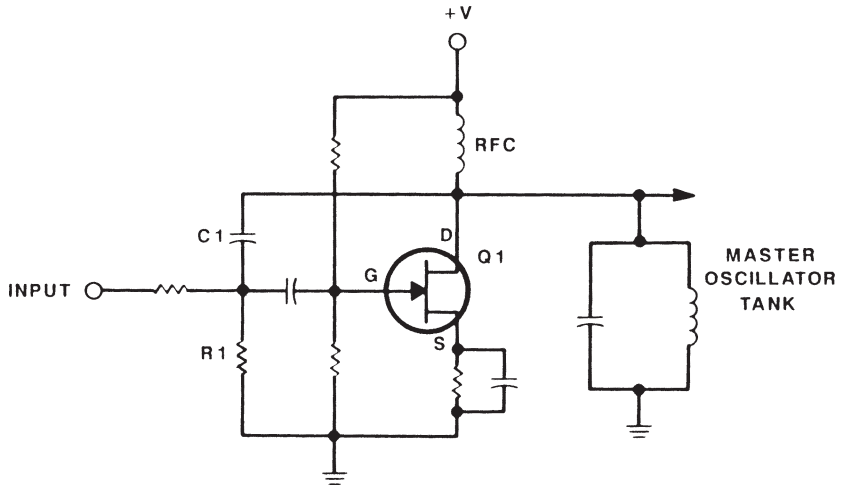


Figure 12-1

A reactance modulator.

Since the FET circuit acts as a part of the oscillator tank capacitance, it controls the master oscillator frequency. When the input is zero, the capacitive reactance is at its normal value, which is determined by the biasing components and C1/R1. When the input goes positive, the FET drain current increases, and the capacitive reactance of the circuit also increases. This adds more capacitance to the oscillator tank; and, as a result, the oscillator frequency is reduced. Likewise, when the input goes negative, the FET current decreases, also decreasing the

capacitive reactance. This in turn increases the master oscillator's operating frequency.

A bipolar transistor, or even a vacuum tube, could be used in place of the FET. The circuit can also be biased to present a variable inductive reactance by reversing the C1/R1 connection. One disadvantage of the reactance modulator is that it cannot easily be used with a crystal-controlled oscillator. Therefore, it requires some form of frequency control circuitry to ensure overall frequency stability. This makes it an expensive and complex circuit, particularly when compared to the varactor modulator discussed next.

VARACTOR MODULATOR

A typical varactor diode modulator is shown in Figure 12-2. It consists of a varactor diode, D1, in a Pierce crystal oscillator circuit. R1 and R2 reverse bias the varactor and set the diode's nominal capacitance value. This, in turn, establishes the oscillator's center frequency. When an input signal is applied, it alternately adds to and subtracts from the varactor's reverse bias voltage. This changes the varactor's capacitance and, therefore, the oscillator's frequency.

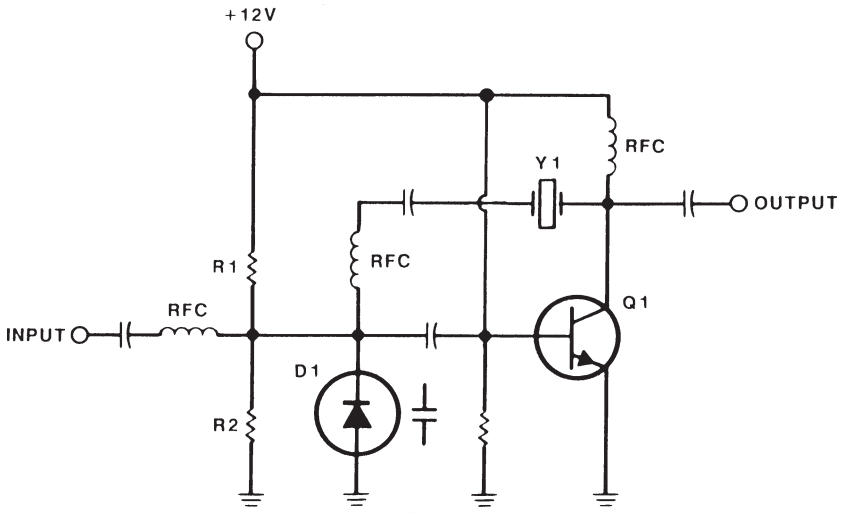


Figure 12-2
A varactor modulated crystal oscillator.

When the signal swings positive, the reverse bias on the varactor increases. This decreases the capacitance of the varactor which, in turn, causes the oscillator frequency to increase. When the signal swings negative, the reverse bias decreases. This increases the capacitance, forcing the oscillator frequency lower.

Thus, the rate of deviation is the same as the input signal. When the amplitude of the signal is increased, the varactor swings through a larger range of capacitance. Consequently, the amount of frequency deviation increases.

The varactor modulator is by far the most popular circuit used to obtain angle modulation. It offers simplicity, reliability, and the stability of a crystal oscillator.

PHASE MODULATOR

Indirect angle modulation is obtained by using a phase modulator, such as that shown in Figure 12-3. The RF input signal is usually from the master oscillator or a buffer stage. This input RF is applied simultaneously to the gate of Q1 and through C1 to the output tank of L1 and C2.

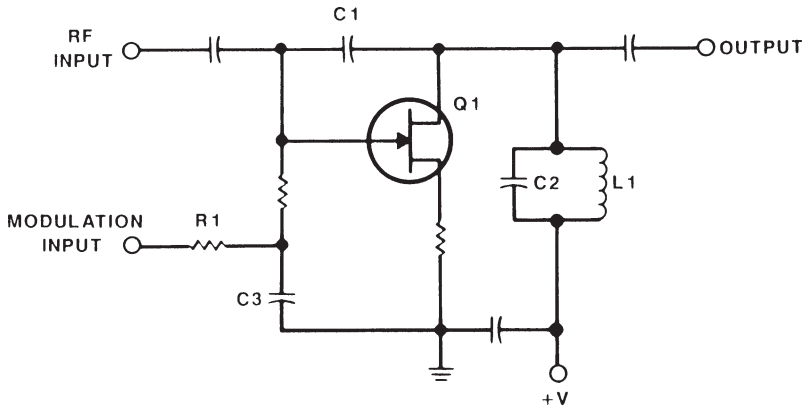


Figure 12-3

An FET phase modulator.

A phasor, or vector, diagram of this circuit is shown in Figure 12-4. Figure 12-4A shows the gate voltage phasor, E_G, and the tank voltage phasor, EL. EL is leading E_G by a given phase angle, θ, due to the phase shift imparted by C1. In addition, a second RF component appears across L1 and C2: the FET output drain voltage, E_D. This voltage leads E by 180° due to the FET's phase inversion. The phase diagram for these components is shown in Figure 12-4B. The output voltage of this circuit, E_O, is therefore the phasor sum of E_D and E_L as is shown in Figure 12-4B.

When a modulating signal is applied to Q1, the gate bias varies and so too does the stage gain. The result is a changing E_D phasor and in turn an E_O phasor that not only changes amplitude, but also phase. This is shown in Figures 12-4B, C, and D. Therefore, the output of the phase modulator is both phase and amplitude modulated. To remove the amplitude variations, the output signal must be amplitude-limited, or clipped. Class C amplifiers can be used to accomplish this. Clipping does not affect the phase modulation, but it does remove the AM component.

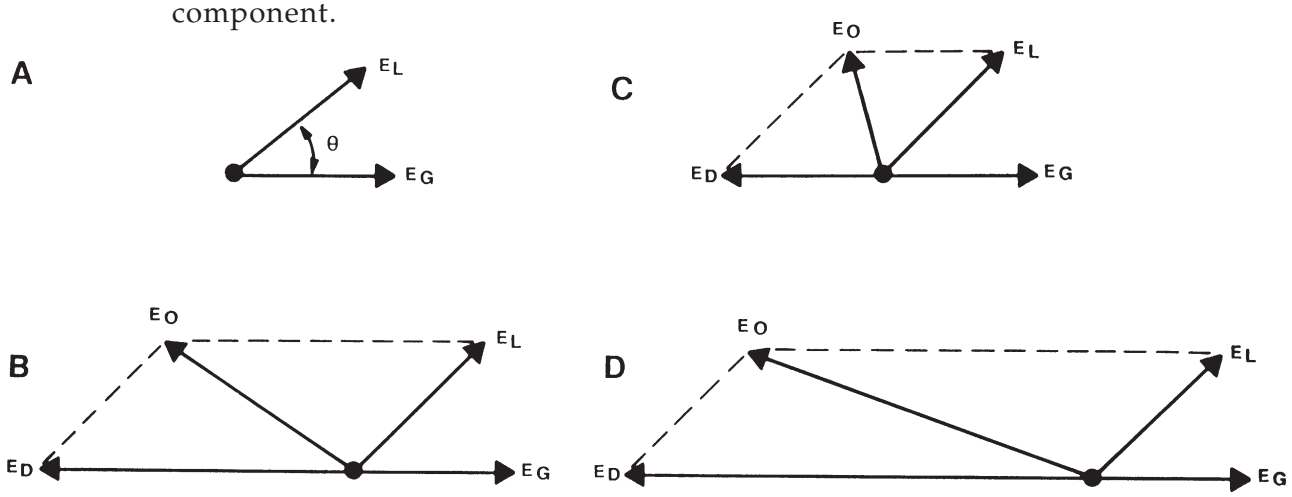


Figure 12-4

Phasor diagrams of the phase modulator.

The advantage of the phase modulator is that the master oscillator is unaffected by modulation. Therefore, it offers increased frequency stability over the direct modulation techniques. It is difficult to obtain wide frequency deviations with a phase modulator. Therefore, its primary uses are in narrow band, two-way FM communications, and data communications, rather than wide band FM broadcasting. You should also recall from Unit 1, that a phase modulator can produce frequency modulation if the input passes through a low pass network before modulation takes place. This is accomplished in Figure 4-3 by the low pass RC network formed by R1 and C3.

A COMPLETE TRANSMITTER

The block diagram of an FM radio transmitter is shown in Figure 12-5. The audio signal is amplified and applied to the FM modulator. The modulator varies the master oscillator frequency according to the frequency and amplitude of the audio signal. The FM carrier is then amplified and transmitted.

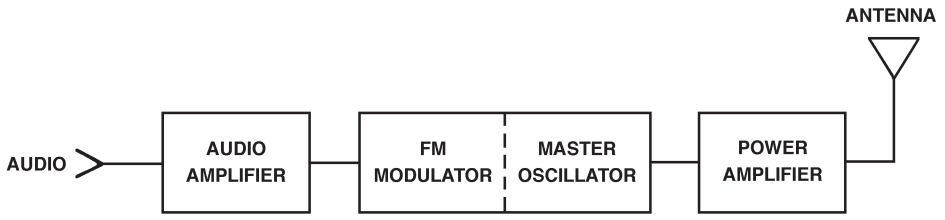


Figure 12-5

Simple FM transmitter.

This type of transmitter has several practical problems at high frequencies. For example, in an FM broadcast transmitter, the output signal is between 88 and 108 MHz. At this high frequency, the master oscillator would have a great deal of difficulty generating a stable frequency carrier. Also, at this high frequency, it is difficult to shift the carrier the required amount without causing distortion.

Many of these problems can be solved by using a different approach. Figure 12-6 shows the block diagram of an FM transmitter that uses a much lower master oscillator frequency.

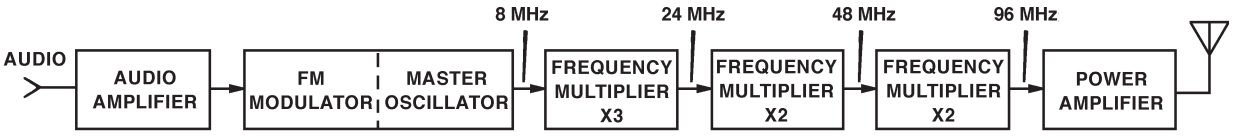


Figure 12-6

An FM broadcast transmitter.

Here, the master oscillator frequency is only 8 MHz. To raise the frequency to 96 MHz, a series of frequency multipliers are used. Remember, a frequency multiplier is nothing more than a nonlinear amplifier whose output is tuned to a harmonic of the input frequency. The most common multipliers are doublers and triplers because efficiency drops when higher multipliers are used. Here, one tripler and two doublers are used to multiply the 8 MHz signal by:

$$3 \times 2 \times 2 = 12$$

Thus, the output frequency is 8 MHz $\times$ 12 = 96 MHz.

The maximum deviation for FM broadcasting is ±75 kHz. Therefore, the output frequency must deviate this amount when an audio signal is applied. However, frequency multipliers increase carrier deviation as well as carrier frequency. For example, suppose the 8 MHz oscillator is varied from 7.9 MHz to 8.1 MHz. After this is multiplied 12 times, the carrier varies from 94.8 MHz to 97.2 MHz. Thus, to obtain ±75 kHz deviation at the output, the master oscillator must be deviated:

$$\pm 75 \text{ kHz} \div 12 = \pm 6.25 \text{ kHz}$$

This can be easily accomplished with a varactor modulator.

One advantage that FM has over AM is that the power amplifier can operate in class C, which is more efficient. This is true, because the FM waveform has a constant amplitude envelope-only the frequency variations need to be reproduced accurately.

Don't forget that FM and PM are used in many more types of communications equipment than radio transmitters. A transmitter has been shown here to give you an idea of how the building blocks are put together. It is also representative, on the block level, of the structure of most FM/PM signal generating devices. But, don't fall into the trap of thinking only about radio! you'll see many more applications before the course is over.

SUMMARY

This chapter introduced you to the circuitry that is used to perform angle modulation. Using these circuits and your knowledge of electronics, you then examined how FM transmitters are built from basic circuit "building blocks."

SELF-TEST REVIEW

1. Name the two types of angle modulation.

   A. _____

   B. _____

2. Name the three different angle modulator circuits.

   A. _____

   B. _____

   C. _____

3. In the FM transmitter shown in Figure 12-7, a master oscillator crystal frequency of _____ MHz is used to obtain the desired 90 MHz output.

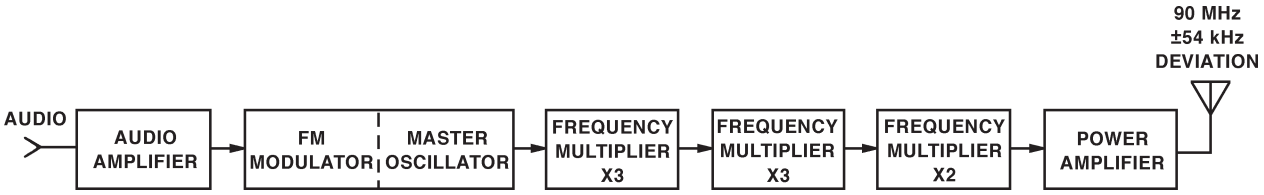


Figure 12-7

Figure for Questions 3 and 4.

4. To obtain 54 kHz deviation at the output of the transmitter shown in Figure 12-7, the master oscillator must be deviated _____ kHz.

5. Using Figure 12-8, complete the block diagram of an angle modulation transmitter using phase modulation.

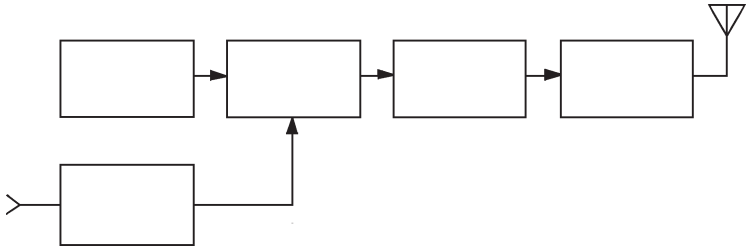


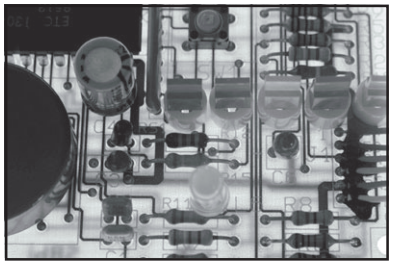
Figure 12-8

Figure for Question 5.

CHAPTER 13 — Demodulators

INTRODUCTION

Now that you know how to create an angle-modulated signal, you need to know how to demodulate it. There are many ways to demodulate an angle-modulated signal. The methods that are easiest to understand are not used very often, but to help you understand the more complex methods, we will discuss the simple ones first. We will then discuss the more complex (but also more popular) circuits.



OBJECTIVES

When you complete this chapter, you will be able to:

- Name five types of FM demodulators, list their advantages and disadvantages, and identify their schematic diagrams.
- State the purpose of a limiter stage.
- Draw the block diagram of a phase-locked loop and explain how it works.

SLOPE DETECTOR

The simplest FM detector is called a slope detector and is shown in Figure 13-1A. The two parallel resonant tank circuits L1/C1 and L2/C2, have an overall response curve like that shown in Figure 13-1B. Notice that the response curve does not peak at the center frequency of the IF signal. Instead, the center frequency falls half way up the side of the curve.

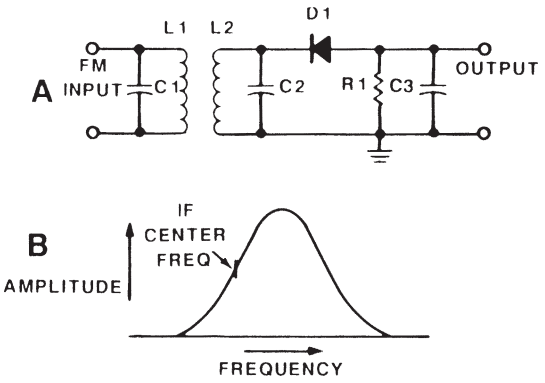


Figure 13-1

The slope detector and its response curve.

When the IF signal is at its center frequency, an average amplitude signal reaches D1. When the IF swings above the center frequency, it approaches the peak of the response curve. Consequently, a higher amplitude signal is applied to the diode. When the IF swings below the center frequency, a low amplitude signal is applied to the diode.

In effect, the tuned circuit changes the frequency modulated IF signal to a signal which varies in amplitude. That is, it changes the FM signal to an AM signal. The diode then detects the AM signal just like the AM detector discussed earlier.

The disadvantage of this technique is that you must also consider the response curve of the IF amplifiers which precede this stage. Having the IF amplifier response curve offset is undesirable, because receiver gain is obviously reduced. Also, acceptable linearity is very difficult to achieve during the FM-to-AM conversion. Consequently, this technique is seldom used in practical applications.

DOUBLE-TUNED DETECTOR

Another FM detector that is easy to understand is shown in Figure 13-2. This one is called a "double-tuned detector". Transformer T1 has a center-tapped secondary. The upper half of the secondary is tuned by C1 to resonate slightly above the center frequency of the IF signal. Likewise, the lower half of the secondary is tuned by C2 to resonate slightly below the center frequency.

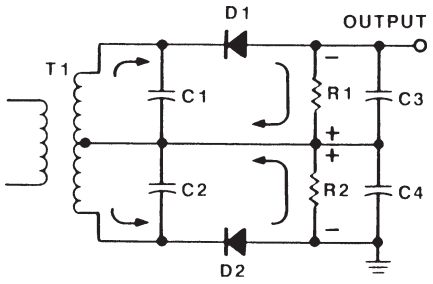


Figure 13-2
The double tuned detector.

When the IF signal is at its center frequency, both halves of the secondary are equally "detuned." D1 and D2 conduct equally. Notice that the voltages developed across R1 and R2 are of opposite polarity and, therefore, tend to cancel. Consequently, if D1 and D2 conduct equally, the output voltage is zero.

Above the center frequency, the signal approaches the resonant frequency of the upper tank circuit. Thus, a larger signal is coupled to D1 than to D2; D1 conducts harder developing a larger voltage across R1 than is developed across R2. Therefore, the resultant is a negative voltage.

When the carrier swings below the center frequency, the signal approaches the resonant frequency of the lower tank circuit. D2 receives a stronger signal than D1; D2 conducts harder, developing a larger signal across R2 than is developed across R1. Since the voltage across R2 is positive with respect to ground, the net output voltage is positive. As you can see, the output signal corresponds to the intelligence contained in the FM input signal.

The primary disadvantage of the double-tuned detector is the precise tuning required for each resonant circuit. If one of the tank circuits is slightly out of tune, the detected audio will be distorted. For this reason, a double-tuned detector is seldom used to demodulate FM audio signals. However, it is used to demodulate frequency-shift

keyed (FSK) signals. These are digital transmissions where a high level, or binary 1, corresponds to one frequency and a low level, or binary 0, corresponds to another frequency, which is shifted slightly from the first. Here, one resonant circuit is tuned to the binary 1 frequency and the other is tuned to the binary 0 frequency. The demodulated output is a pulse train which can be coupled to a digital computer or other peripheral device. Usually, the FSK signal is at audio frequencies, for example 1070 Hz and 1270 Hz, and, therefore, the demodulator is also tuned to these frequencies. By using audio frequencies, data signals can be transmitted over the telephone lines.

FOSTER-SEELEY DISCRIMINATOR

A more popular FM detector is shown in Figure 13-3. This one is called the "Foster-Seeley discriminator."

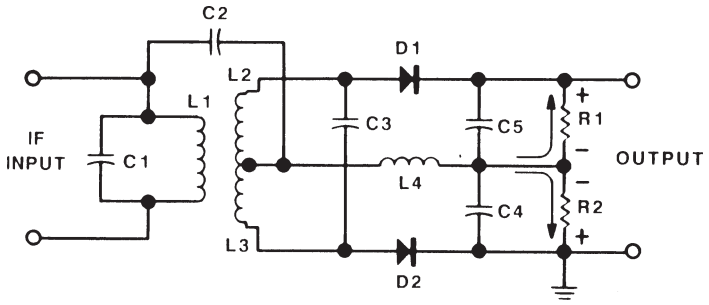


Figure 13-3

The Foster-Seeley discriminator.

The input to the circuit is an IF signal which is varying at the modulation rate. The output is the detect signal. The two diodes and their associated RC networks operate similar to their counterparts in the double-tuned detector discussed earlier. That is, when both diodes conduct, equal but opposite polarity voltages are developed across R1 and R2. The two voltages cancel, and the output is 0 volts. However, if D1 conducts harder, the output is positive. By the same token, the output is negative when D2 conducts harder. The audio signal can be recovered from the IF signal if:

1. Both diodes conduct equally at the center frequency.

2. D1 conducts harder above the center frequency.

3. D2 conducts harder below the center frequency.

What determines how much each diode conducts?

By transformer action, L1 couples the IF signal to L2 and L3. As connected, L2 and L3 act as a center-tapped secondary. Thus, the voltage developed across L2 (E_{L2}) is 180° out of phase with the voltage developed across L3 (E_{L3}). E_{L2} controls the conduction of D1, while E_{L3} controls the conduction of D2. Keep in mind that these two voltages are equal in amplitude, but 180° out of phase.

The IF signal is also capacitive coupled through C2 to L4. A voltage, which we will call EL4, is developed across L4. E_{L4} controls the conduction of both D1 and D2.

The circuit is arranged so that E_{L4} is 90° out of phase with both E_{L2} and E_{L3}. However, as you will see, this is so only when the IF signal is at its center frequency. At the center frequency, E_{L4} leads E_{L3} by 90° but lags behind E_{L2} by 90°.

The amount that diode D1 conducts is determined by E_{L2} and E_{L4}, while the amount that D2 conducts is determined by E_{L3} and E_{L4}. We have seen that E_{L2} and E_{L3} are equal in amplitude but are 180° out of phase. At the center frequency, E_{L4} is 90° out of phase with both E_{L2} and E_{L3}. Thus, at the center frequency, E_{L4} adds to both of these signals equally. Consequently, D1 and D2 conduct equally, and the output voltage is 0.

The parallel resonant circuit is resonant at the center frequency, where X_L exactly cancels X_C, and the resonant circuit acts resistive. However, above resonance, X_L is larger than X_C. Thus, there is a net reactance that shifts the phase of E_{L2} and E_{L3}. E_{L2} is shifted more in phase with E_{L4}. Since E_{L3} is always 180° out of phase with E_{L2}, E_{L3} is shifted more out of phase with E_{L4}. That is E_{L4} tends to add to E_{L2} but tends to subtract from E_{L3}. Thus D1 conducts harder than D2. The net result is that the output swings positive each time the IF signal swings above the center frequency.

Below resonance X_C is larger than X_L. The net reactance shifts the phase of E_{L2} and E_{L3} in the opposite direction. This time, E_{L3} is shifted more in phase with E_{L4}. Consequently, D2 conducts harder, producing a net negative output voltage. Thus, the output swings negative each time the IF signal swings below the center frequency.

If the modulating signal is a 1000 Hz tone, the IF swings above and below the center frequency 1000 times each second. The discriminator produces an output sine wave that swings positive then negative at

the same rate. Thus, the discriminator recovers the modulating signal. One of the disadvantages of the Foster-Seeley discriminator, as well as the other FM detectors previously discussed, is that it is also sensitive to amplitude changes. That is, any noise spikes present on the FM wave are detected and passed on to the audio output as noise, or "static." However, since an FM wave does not have an envelope that must be preserved, we can use a special circuit to clip off both the positive and negative peaks of the wave. The frequency variations remain but the majority of noise impulses are removed. This circuit is called a "limiter", and it immediately precedes the FM demodulator.

RATIO DETECTOR

A further improvement over the Foster-Seeley discriminator is the ratio detector, shown in Figure 13-4. This circuit is one of the most popular types of FM detectors because is provides its own limiting action, and it requires no preceding limiter stage.

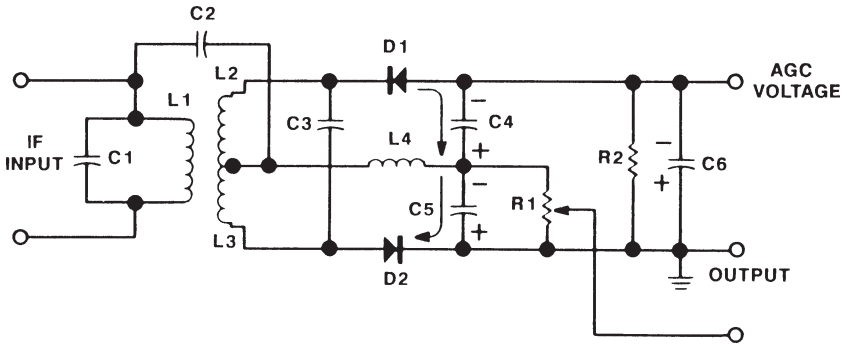


Figure 13-4

The ratio detector.

The ratio detector circuit looks somewhat similar to the Foster-Seeley discriminator. However, closer examination shows that diode D1 is reversed, and the output configuration is different.

With D1 reversed, the two diodes are in series across the entire secondary. Conduction of the two diodes is controlled by the same factors as in the Foster-Seeley discriminator. At the center frequency, the two diodes conduct equally. In this circuit, the voltages build up across series capacitors C4 and C5. Recall that in the discriminator, the diodes produce opposing voltages.

The key to the unique operation of the ratio detector is C6. C6 is a large value capacitor. This capacitor charges to a voltage which is proportional to the average received signal strength. C6 is large enough to hold the voltage across R2 constant. It also holds the voltage constant across the series combination of C4 and C5. It does this even if there are momentary amplitude variations (noise). This is the reason the ratio detector is relatively insensitive to noise.

C6 also provides a convenient AGC voltage, since the voltage across it is directly proportional to the input signal amplitude.

The voltage across C5, plus the voltage across C4, must always equal the voltage across C6. At the center frequency, the two diodes conduct equally, and the voltages across the two capacitors are equal. A sample of the voltage across C5 is tapped from R1. This is the ratio detector output. The detector output is at some negative DC level.

As in the discriminator, D1 and D2 alternately conduct harder as the IF signal swings above and below the center frequency. When the diodes are not conducting equally, the difference current flows through L4. When D1 conducts harder, the voltage across C4 exceeds that across C5. However, since the sum of these two voltages remains constant, the voltage across C5 must decrease. Consequently, the voltage at the output must also decrease.

When D2 conducts harder, a higher voltage develops across C5. Therefore, the output voltage also increases. As you can see, the output voltage swings in step with the frequency changes in the IF signal. That is, when the frequency increases, the output voltage increases. And, when the frequency decreases, the output voltage decreases.

The primary advantage of the ratio detector is its self-limiting action. It also provides AGC voltage.

All of the preceding demodulators depend on the characteristics of tuned circuits to recover the modulating signal. This is not often desirable, as such circuits need frequent alignment, and are notoriously nonlinear. It is preferable to use some sort of system that does not need tuning (or at least not much) and which recovers the modulating waveform more or less directly. This is done using feedback demodulators, of which the phase-locked loop, or (PLL), is the most commonly used.

PHASE-LOCKED LOOP DEMODULATOR

It has always been possible to use phase-locked loops as FM demodulators, but they were not practical until integrated circuits were developed. A complete phase-locked loop, like that shown in Figure 13-5, can be purchased as a single integrated circuit. This makes the PLL simple and inexpensive to use. The basic loop of Figure 13-5 consists of a phase detector, low pass filter, and voltage-controlled oscillator. The voltage-controlled oscillator, or VCO, operates at the input frequency, in this case, the intermediate frequency. The phase detector compares the VCO signal and the input signal. It then develops an error voltage that is proportional to the amount and direction of the phase difference between these two signals. This error signal is then coupled to the low pass filter.

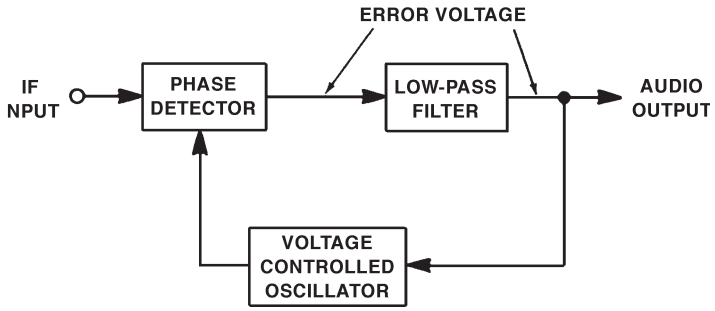


Figure 13-5

A phase-locked loop demodulator.

The key to this circuit is the phase detector, of which there are several types. Remember that a phase detector "sees" a phase difference in either of these cases:

A. The two signals applied are of the same frequency, but differ in relative phase.

B. The two signals applied differ in frequency, in which case, they differ in phase by an amount proportional to the difference in their frequencies.

The error voltage from the low pass filter is used to control the VCO. For example, if the input frequency deviates upward, an error voltage is generated by the phase detector. This voltage is filtered and applied to the VCO. The error voltage causes the VCO to move upward in

frequency in an exact lock with the input frequency. When the input signal is frequency modulated, the VCO tracks the FM deviation and, as a result, the error voltage is an exact duplicate of the modulating information.

Figure 13-6 shows a typical integrated circuit PLL. In this circuit, R1 and C1 are used to set the VCO's operating frequency range. C2 and the internal 3.6 kΩ resistor form an RC low pass filter. The only difference between this circuit and the block diagram of Figure 13-5, is the DC amplifier, which increases the error signal voltage.

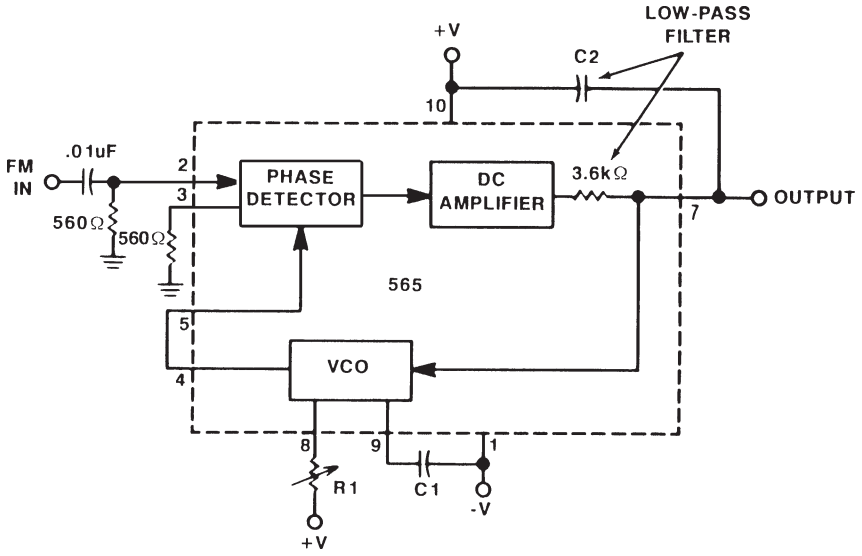


Figure 13-6

An IC phase-locked loop.

The primary advantage of the phase-locked loop is its excellent performance at low cost, and with a minimum of components. PLLs are easily aligned, and perform better in noise than do discriminators. As a result, they are widely used as FM and PM demodulators.

Note that the error signal to which the PLL responds depends upon the phase difference between the applied signal and the VCO output. That is, the error signal actually represents the voltage required to produce a phase modulated signal of the same waveform as the one presented for detection; the error voltage waveform looks just like the modulating waveform used to generate PM. Because of this, phase locked loop modulators can be used to demodulate phase modulated signals, which are frequently used in data communication systems.

SUMMARY

In this chapter you learned about five different types of FM detectors. You learned that the slope detector is the simplest circuit to understand, but is least used because it places restrictions on its preceding stages.

Of the detectors discussed, the phase-locked loop demodulator is the most used. It performs well, is low cost, uses a small number of components.

SELF-TEST REVIEW

1. Name 5 types of FM demodulators.

   A. _____

   B. _____

   C. _____

   D. _____

   E. _____

2. What are the advantages and disadvantages of each of the FM demodulators named in Question 1?

   A. _____
   _____

   B. _____
   _____

   C. _____
   _____

   D. _____
   _____

   E. _____
   _____

3. What type of demodulator is shown in Figure 13-7? Which terminal is used for the audio output? Which terminal is used for AGC voltage? _____

_____

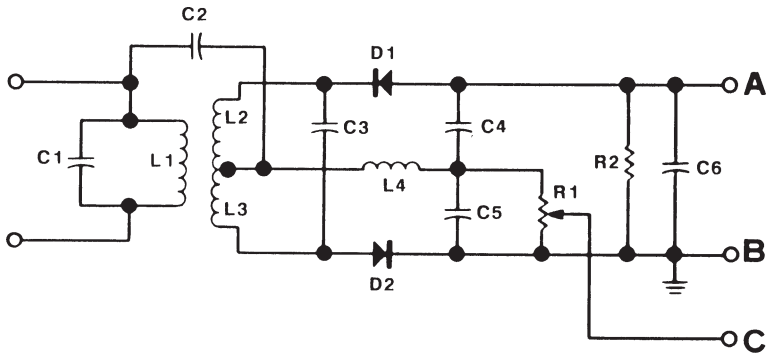


Figure 13-7

Figure for Question 3.

4. What circuit is shown in Figure 13-8? Which terminal is used for the audio output? Which terminal is used for AGC voltage?

_____

_____

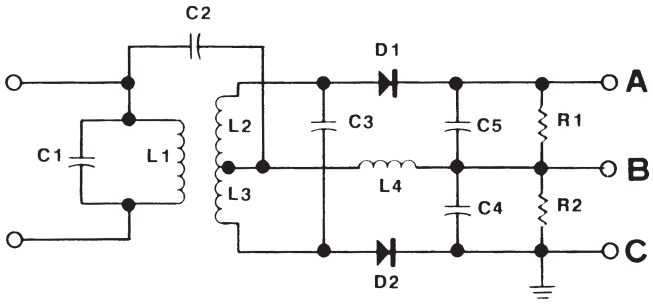


Figure 13-8

Figure for Question 4.

5. What is the purpose of a limiter? _____

_____

_____

6. What FM demodulator does not require a limiter?

_____

_____

7. Using Figure 13-9, complete the block diagram of a phase-locked loop, and label the demodulated output.

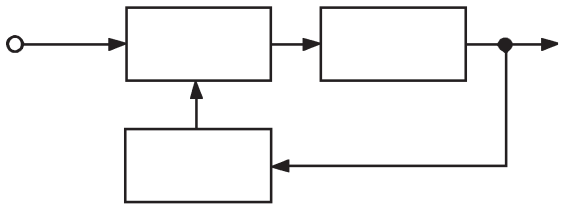


Figure 13-9

Figure for Question 7.

CHAPTER 14 — Receivers

INTRODUCTION

To understand how angle modulated signals are received and detected, we are going to look at an FM receiver in some detail. As with the transmitter in an earlier section, however, don't limit your thinking to radio. This is just an illustration of one application of angle modulation; however, the principles shown apply to many types of receiving equipment.

An FM receiver is actually very similar to the AM superheterodyne receiver discussed previously. That is, it converts the received RF signal to a lower intermediate frequency, where the majority of gain and selectivity are obtained. The only difference is it uses an FM detector and possibly a limiter stage.

OBJECTIVES

When you complete this chapter, you will be able to:

- State the purpose of a limiter stage.
- Explain the purpose of AFC.
- Label the blocks of a typical FM tuner.

FM BROADCAST RECEIVER

The block diagram of typical FM broadcast receiver is shown in Figure 14-1. Notice the similarity between this receiver and an AM receiver. The primary differences are the intermediate frequency, the limiter, and the FM detector. The IF of 10.7 MHz is chosen to provide good image rejection; and also, to provide the wider bandwidth required for FM broadcast signals. In FM broadcasting the maximum transmitter deviation is 75 kHz. Therefore, the receiver bandwidth must be at least 150 kHz, and is usually 180 to 200 kHz. The IF of 10.7 MHz allows this, while maintaining acceptable IF gain.

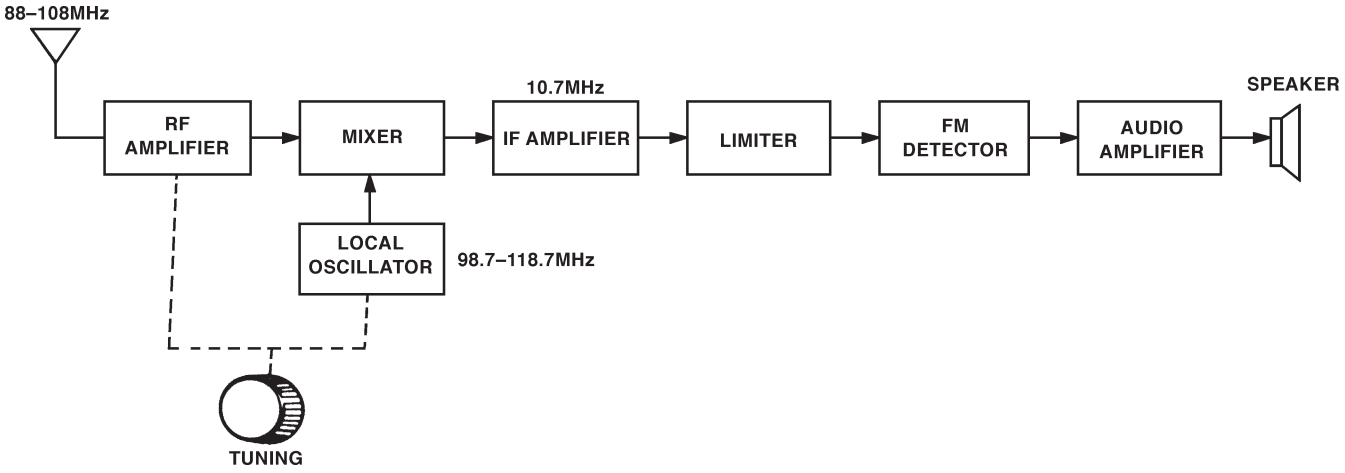


Figure 14-1

An FM broadcast receiver.

Let's follow the signal through the receiver. Assume that the receiver is tuned to 100 MHz. When the tuning indicator is set to this frequency, the RF amplifier is tuned to 100 MHz, while the local oscillator is tuned to 110.7 MHz. The RF and local oscillator signals heterodyne in the mixer, forming a difference signal of 110.7 – 100 = 10.7 MHz. This signal is passed to the IF amplifier where it receives most of its amplification. The IF amplifier also reduces the received bandwidth sufficiently to select the correct signal while rejecting all others.

If the received carrier is deviating 50 kHz, and the local oscillator frequency is constant, the 10.7 MHz IF signal also deviates 50 kHz. Thus, the original modulation is converted to the intermediate frequency without distortion.

The next stage in the block diagram is the limiter. As mentioned earlier, this circuit is used to remove any amplitude variations, such as noise, from the FM signal. It does this by saturating on the positive half cycle and going into cutoff on the negative half cycle. It, therefore, clips or cuts off both the positive and negative peaks of the waveform and, along with them, most of the noise. This limiter action is shown in Figure 14-2. Note also that the frequency variations of the original modulation are unaffected by the amplitude clipping.

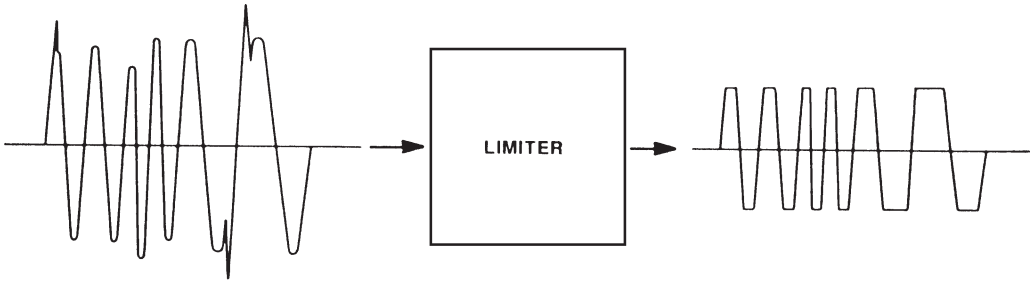


Figure 14-2

The limiter removes noise spikes and other amplitude variations.

Note that the limiter stage must be overdriven to a certain extent. That is, if the input amplitude is not high enough to drive the limiter into cutoff and saturation, noise passes on to the FM detector, and eventually the audio output. In fact, one FM receiver specification refers to the "quieting" of the audio output. This is the sensitivity of an FM receiver. It specifies the unmodulated input signal required to reduce the audio noise output by 20 dB. The only way to reduce the noise output is to amplify the input signal sufficiently to overdrive the limiter. The limiter clips the noise pulses from the wave and, in the absence of modulation, provides a "quiet" audio output.

When an FM receiver is operated at VHF, as it is in FM broadcasting, the local oscillator must be extremely stable. For example, a frequency drift of just 0.1% at 100 MHz causes the IF to drift 100 kHz. This results in serious distortion of the demodulated wave. One way to prevent drift is to use a crystal-controlled local oscillator. Crystals are used in communications, where only a few frequencies are required. But crystals are impractical for broadcast reception since you want to select many different frequencies without buying a new crystal each time. In FM broadcasting, an automatic frequency control, or AFC, circuit is used to "lock" the local oscillator onto the incoming signal. Figure 14-3 shows the block diagram of an FM receiver that uses AFC.

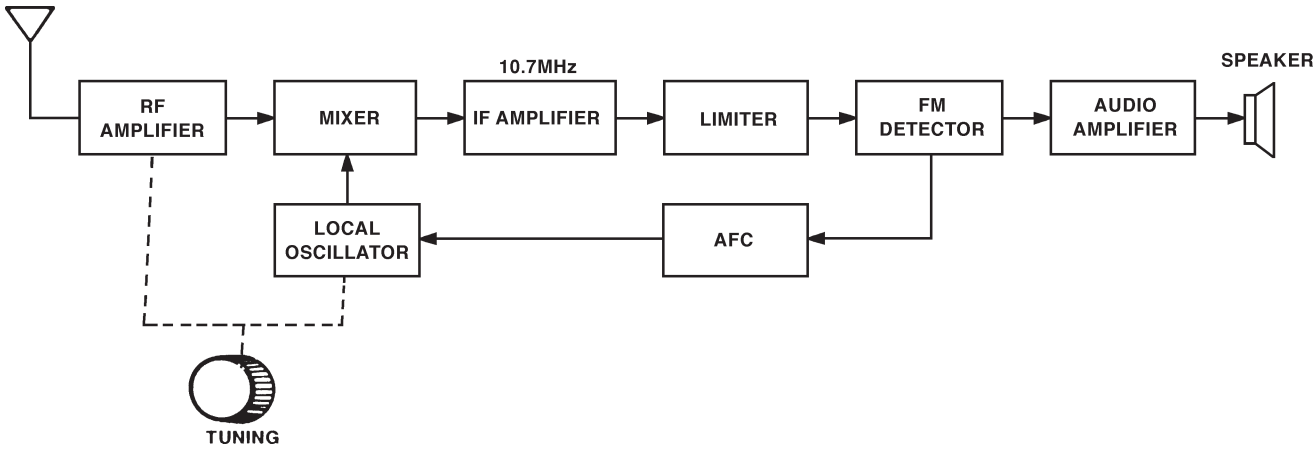


Figure 14-3

Automatic frequency control in an FM receiver.

The basic principle of AFC is to use a varactor, or a similar device, as part of the local oscillator tank. The varactor's capacitance is controlled by an "error" voltage that is produced when the oscillator goes off frequency. This error voltage is obtained from the output of the discriminator, or ratio detector. Here, if the oscillator is on frequency, the output is zero. However, if the frequency drifts to one side or the other, an error voltage is produced. The polarity of this voltage indicates the direction of drift and the amplitude indicates the magnitude of frequency drift.

It seems that it would be a simple matter to use this error voltage to control the local oscillator. However, the output of the discriminator also contains the demodulated audio wave. Therefore, before this voltage can be used to control the oscillator, the audio component must be removed; otherwise, the AFC action would cancel the frequency modulation. The audio can be removed by using a low pass filter. This removes the audio while allowing the DC and the very low frequency error signal to pass on to the varactor.

Figure 14-4 shows a simplified AFC circuit. Q1, and its associated circuitry, form the local oscillator. C2 is the receiver's main tuning control, while D1 is a varactor which is in parallel with the oscillator tank. In this circuit, the output of the FM detector is coupled to both the audio amplifiers and the AFC circuit. R1 and C1 form a low pass filter which removes the audio component from the signal going to the AFC. Then this DC error voltage either adds to, or subtracts from, the varactor bias provided by R2 and R3. R4 is used to isolate the AGC filter from the oscillator. As the varactor's capacitance changes, so does the local oscillator frequency.

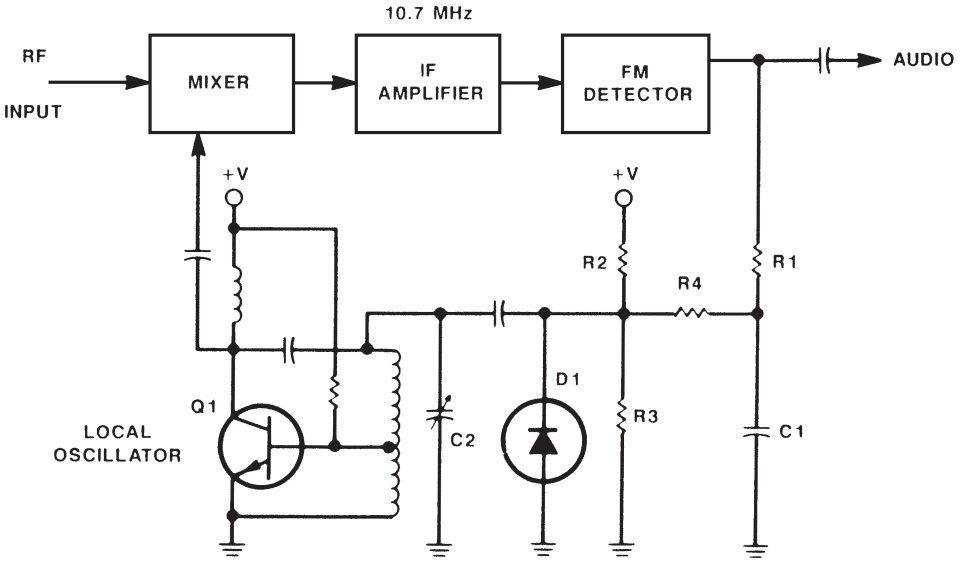


Figure 14-4

Simplified AFC circuit.

Notice that the AFC circuit in Figure 14-4 is actually a variation of the phase-locked loop. In this case, the FM detector is the phase detector, R1 and C1 are the filter, Q1 and D1 form the VCO, and the mixer is used to convert the two VHF signals (RF input and local oscillator) to a much lower frequency and to provide the needed 10.7 MHz offset. The filter parameters were chosen so that the feedback loop responds only to slow frequency changes, such as would be produced by oscillator drift, and not to the rapid frequency variations due to the modulation.

AM/FM BROADCAST TUNER

The schematic diagram of a typical AM/FM tuner, such as that shown in Figure 14-5, shows you a practical application of the circuits we've discussed so far. Remove "Schematic Diagram #2" from the schematics packet included with this course. The signal path is marked by a red dashed line. Let's follow it through to the audio output.

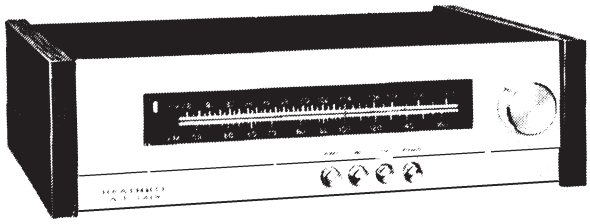


Figure 14-5
The Heathkit AJ-1219 AM/FM Tuner.

From the antenna input, which can be either 300Ω balanced or 75Ω unbalanced, the signal is coupled to the RF amplifier, Q301. This FET amplifier has tuned circuits on both its input and output. This improves its selectivity and gain, but lowers its stability. For that reason, L301 and C306 couple energy from the drain tuned circuit back to the gate. The phasing of this transformer coupling provides degenerative feedback and, therefore, stabilizes the amplifier.

Once the RF input is selected and amplified, it is applied to the mixer, Q302. The signal from the local oscillator, Q303, is also applied to the mixer, as shown by the red dotted line. The output of the mixer is the difference signal of 10.7 MHz, the intermediate frequency. This IF signal is coupled through T302 to the 10.7 MHz ceramic filter, F1. The ceramic filter, used to improve receiver selectivity, operates in the same manner as a crystal filter, since it exhibits piezoelectric effects similar to quartz. In addition, the ceramic filter is smaller, and cheaper, than a comparable crystal filter.

The IF amplifier, IC1, is an MC1357 integrated circuit that is a combination IF amplifier, limiter, and FM detector. However it is only connected as an IF amplifier and limiter. In this configuration, it can provide up to 60 dB gain.

IC1 has two output connections: pins 10 and 14. The output from pin 10 is applied to Q6, which is the AGC amplifier. The amplified IF signal from Q6 is rectified by the voltage double circuit of D1 and D2 and C29 and C27. After further filtering by R21 and C25, the resulting negative DC voltage is directly proportional to the input signal level. This voltage is then used to control the gain of the RF amplifier by changing the gate bias. This is shown on the schematic diagram by the dotted green line.

The output from pin 14 of IC1 is applied to the second 10.7 MHz ceramic filter, F2. This provides additional IF selectivity. The output of F2 is then coupled to IC2, which is also an MC1357 integrated circuit. It too provides IF amplification and limiting.

In addition, it is configured to operate as an FM detector. The output of IC2, at pin 1, is the demodulated audio signal. This audio signal is applied to Q3, which is an audio amplifier.

The demodulated audio signal is also coupled to the AFC circuit as indicated by the dashed green line. In the AFC circuit, the audio component is removed by the low pass filter comprised of R28 and C6. The remaining DC voltage, which is directly proportional to the frequency drift, is used to control the bias of varactor D301. Steady-state bias for D301 is applied by R308 and R309 to the cathode, and R22 and R4 for its anode. Potentiometer R4 is used to set the local oscillator for correct dial calibration. The AFC voltage either adds to or subtracts from D301's anode bias voltage, to provide automatic frequency control.

In FM "Stereo" broadcasting, the output of the demodulator contains frequency components as high as 53 kHz. This is due to the multiplexing techniques used to transmit both the right and left audio signals. This wide range of frequencies is amplified by Q3 and applied to IC3, which is the stereo demodulator. This circuit separates the demodulated audio into the right and left channels. In addition, it provides a signal at pin 6 to indicate when a stereo signal is being received. The signal is used to turn on Q5 and the stereo lamp. FM stereo broadcasting is described in detail in later.

Q8 and Q7 are audio amplifiers for the right and left channels respectively. L4 and L5 and L6 and L7 form low pass filters and traps to prevent any signals above 15 kHz from appearing at the audio output. Since this is a tuner, there are no audio power amplifiers. Instead, a separate power amplifier must be used to drive loudspeakers. You might like to look over the AM section of the tuner as a review of a previous chapter.

SUMMARY

In this chapter you analyzed a broadcast receiver to see how practical angle modulation equipment is built. You also learned that an FM receiver is similar to an AM receiver, except it uses an FM detector and maybe a limiter stage.

SELF-TEST REVIEW

1. Why is 10.7 MHz the standard intermediate frequency in FM receivers? _____

   _____

2. How can a limiter remove noise from an FM signal without causing distortion? _____

   _____

3. What is the purpose of AFC and why is it needed? _____

   _____

   _____

4. Use Schematic Diagram #2 as a reference to complete the block diagram of the FM tuner provided in Figure 14-6.

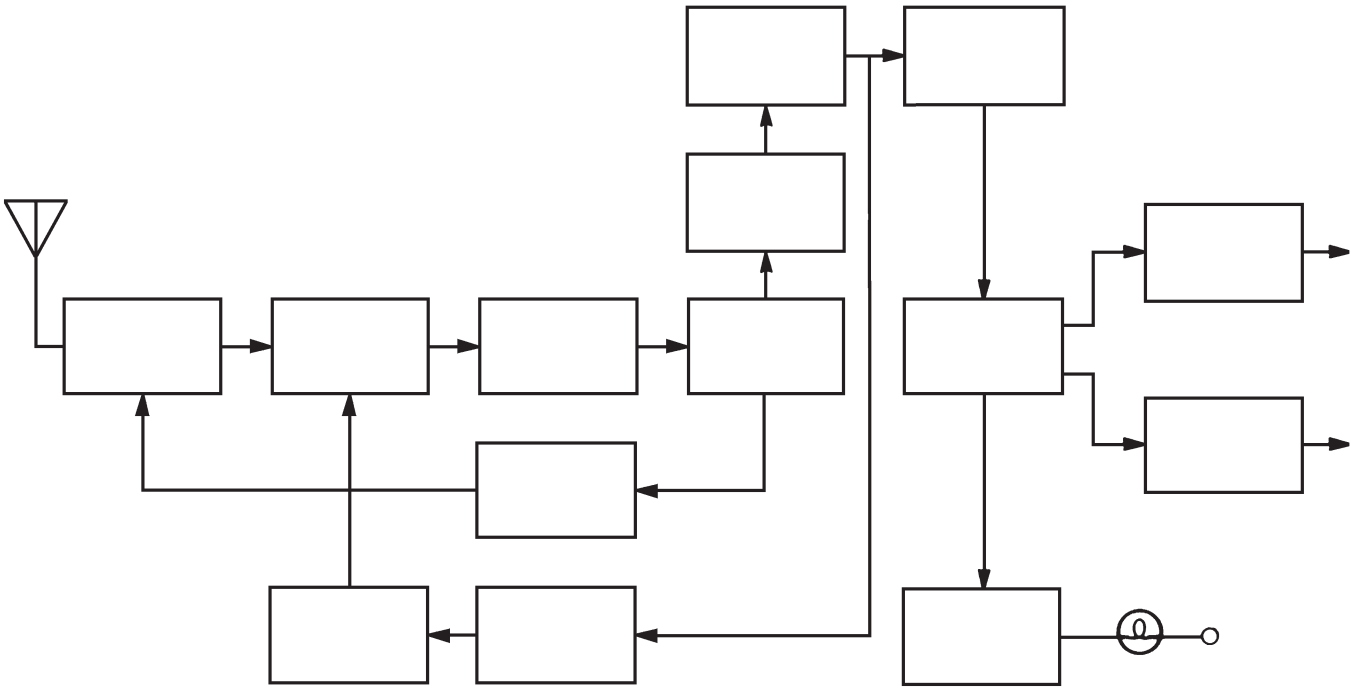


Figure 14-6

Figure for Question 4.

CHAPTER 15 — FM Transceivers

INTRODUCTION

In two-way radio communications, the transmitter and receiver are often combined into a single unit called a "transceiver."

The hand-held VHF transceiver shown in Figure 15-1 contains everything needed for communication. It has both transmitter and receiver; microphone, loudspeaker, antenna, battery power supply, and even a key pad for dialing a telephone through a special base station. In the following text, we will analyze the schematic diagram of a similar transceiver. You will become familiar with some additional circuits, such as "squelch" circuits, often found in practical equipment.

Figure 15-1
A hand-held FM transceiver.

OBJECTIVES

When you complete this chapter, you will be able to:

- Explain the purpose of a squelch circuit.
- Determine the crystal frequency that is required to receive a given frequency when you know the IF frequencies.
- Determine the amount of deviation needed at the crystal oscillator to obtain a given deviation at the output of a transmitter.
- Draw the block diagram of a transmitter.

SCHEMATIC ANALYSIS

Remove "Schematic Diagram #3" from the packet of schematics included with this course. We'll analyze the receiver first and then the transmitter. The receiver signal path is marked by a red dashed line.

The input RF signal appears at the antenna jack and passes through L41 and its associated capacitors. They form a low pass filter that reduces harmonic radiation by the transmitter and protects the receiver from overload by very strong UHF signals. The signal then passes through diodes D6 and D9. In the receive mode, these diodes are forward biased by the receiver's positive supply voltage through R5 and R125. Since they are forward biased, the RF signal passes through unattenuated. However D11, which leads to the transmitter, is reverse biased and it prevents signals from entering the transmitter.

From the diode switches, the signal passes to Q1, the first RF amplifier. It is then coupled to Q2, an additional RF amplifier. Together these two amplifiers provide 28 dB of power gain. They also provide sufficient selectivity to limit reception to a 2 MHz bandwidth. In this case, the frequencies of interest are the amateur radio band from 144 MHz to 148 MHz. The RF amplifiers must be tuned to the desired 2 MHz portion of this band.

The output of the RF amplifiers is applied to gate 1 of mixer Q3. The local oscillator signal, the red dotted line, is applied to gate 2. These two signals are mixed and their difference signal is applied to crystal filters FL1 and FL2. These filters determine the IF selectivity and, therefore, the receiver's bandwidth.

Let's follow the local oscillator signal path. Q9 is the crystal oscillator. Switch SW202 can be set to select any one of 8 different crystals. D3 is a varactor which is used only during transmit to provide frequency modulation. This is direct modulation. The crystal shown at Y1 oscillates at 15.1377 MHz. Not only is Q9 the oscillator, but it also provides a frequency multiplication of 3. This is because the collector load of Q9, L22, and C64, is resonant at the third harmonic of the crystal frequency. Therefore, the output of Q9 is $15.1377 \times 3 = 45.4131$ MHz. This signal is then coupled to a 45 MHz bandpass filter formed by C66, C67, C68, and L23. Q10 amplifies the 45 MHz signal and provides a further multiplication by 3 through its output tank of L24 and C71. Thus, the output of Q10 is $45.4131 \times 3 = 136.2393$ MHz. Q11 provides further amplification, filtering for the 136 MHz signal, and buffering for the oscillator chain. When the local oscillator signal is applied to mixer Q3, the RF input signal must be 136.2393 MHz + 10.7 MHz = 146.9393 MHz to provide a difference signal of 10.7 MHz that is

amplified by the IF section. The crystal oscillator's frequency can be shifted slightly by coil L13 to set the received and transmitted frequencies to exactly the right channel (146.940 MHz).

The output of IF amplifier Q4 is applied to IC1. This integrated circuit is an IF amplifier, FM detector, and audio amplifier. Crystal filter FL3 forms part of the FM detector circuit.

The audio output from pin 6 of IC1 is applied through the volume control to pin 7 of IC2, the audio power amplifier. IC2 is used to drive the speaker, or earphones, if used.

There is one more part of the receiver which we must discuss. This is the squelch circuit. The squelch circuit is used to "cutoff," or "squelch," the audio amplifier when no signal is present. If the audio amplifier were not cut off during no-signal conditions, the noise present would be amplified and applied to the speaker. This can be extremely annoying. The squelch circuit monitors the FM detector. If no signal is present, it cuts off the audio amplifier, and there is no output. If a signal is present, the squelch circuit turns on the audio amplifier and you can hear the signal through the loudspeaker. Let's analyze this particular squelch circuit in detail.

The path of the squelch signal is shown by the solid red line. The output from pin 13 of IC1 is a voltage level that identifies the presence or absence of a signal. That voltage is applied to Q5 which compares it to the voltage coming from the squelch control. The collector voltage of Q5 is applied to IC3B which is wired as a comparator. At Q5, if the signal level voltage at the emitter exceeds the squelch voltage at the base, the transistor is cut off and the collector voltage is high. This high positive voltage causes comparator IC3B to switch, giving a high voltage at pin 7, its output. This high voltage cuts off Q6 and prevents a "turnoff" voltage from being applied to pin 5 of audio amplifier IC2. Therefore, as long as the received signal level is above the squelch control setting, the audio amplifier is turned on and operating.

However, if the input signal is very low or nonexistent, Q5 is turned on, and low voltage is applied to comparator IC3B. This causes the comparator output to go low and, therefore, turn on Q6. With Q6 turned on, a positive voltage is applied to pin 5 of IC2, turning off the audio amplifier.

Connected to the squelch circuit is a special "battery saver" circuit. D5, at the output of IC3B, switches this circuit on and off. IC3A forms an 8 Hz astable multivibrator with the output waveform as shown. When a received signal is present, the comparator output is high. D5

is forward biased, placing a high voltage on pin 2 of IC3A. This shuts off the multivibrator, and, as a result, the battery saver circuit. However, when no input signal is present, the comparator's output is low. This reverse biases D5, which allows the astable multivibrator to operate. The output waveform is used to key Q7 on and off. Since Q7 controls the power supply voltage to the receiver, it is keyed on eight times each second and remains on for only a few milliseconds. This allows the receiver to monitor the input while consuming the least amount of battery power. When a signal is detected, the squelch circuit turns off the astable multivibrator and applies a low voltage to Q7. This holds Q7 on and, thus, applies continuous power to the receiver.

Power to the transmitter is off unless the push-to-talk button is depressed, which turns on Q8 and applies power to the transmitter. At this time, several diodes are forward or reverse biased to control several circuit functions. D7 is forward biased and shuts off receiver audio amplifier IC2. D8 is forward biased, which shuts off the battery saver circuit. This turns off Q7 and, therefore, removes power from the receiver. D11 is forward biased, which connects the antenna to the transmitter. D6 and D9 are reverse biased, disconnecting the receiver from the antenna. D1 is reverse biased, disconnecting the local oscillator power from the receiver voltage line or the transmitter line. D2 is forward biased, connecting the local oscillator power from the transmitter voltage line.

The input audio to the transmitter starts at the microphone. This is shown by the solid green line. IC4A provides voltage gain, while IC4B provides both gain and limiting. Above approximately 4 V peak-to-peak, IC4B clips the signal to limit the maximum transmitter deviation. Output waveforms for both amplifiers are shown on the schematic. The output of IC4B is applied to the low pass filter formed by R86, R87, C85, and C86. The filter removes any high frequency components, or harmonics, from the clipped audio signal. R88 is the deviation control. It sets the maximum transmitter deviation. The modulator output is applied to varactor D3. The varactor's changing capacitance frequency-modulates crystal-controlled oscillator Q9.

You can follow the transmitter signal through the local oscillator to transmitter mixer Q13 by the dashed green line. At the transmit mixer, the output of the offset oscillator is mixed with the local oscillator signal. In simplex operation, which means transmission and reception are on the same frequency, the offset signal is 10.7 MHz. Therefore, the output of the transmit mixer is 10.7 MHz + 136.2393 MHz =146.9393 MHz, which is exactly the same as the received signal.

The offset oscillator can also generate an 11.3 MHz signal or a 10.1 MHz signal. These outputs provide transmitter frequencies, respectively, 600 kHz above or 600 kHz below the receiver frequency. This is done to allow operation with repeater stations whose input and output frequencies are almost always separated by 600 kHz. A repeater station is used to extend the communications range of handheld and low-power transceivers such as the one we are describing. It can receive an input signal and retransmit it at a different frequency, with high power, from a central or more favorable location.

The output of the transmit mixer is applied to a bandpass filter comprised of L31, C96, C97, L32, C99, and C101. This filter insures that the correct mixing product is selected. Transmit predriver Q14 and amplifier Q15 provide additional gain and filtering before the signal reaches driver Q16. The input voltage to Q16 can be varied by R114. This adjustment sets the input and output power levels of the transmitter.

The output of the driver is applied to power amplifier Q17. Its output is approximately 2.5 watts. This signal is coupled to the antenna through a low pass filter and wave trap. C123, C124, L41, and C126 form a low pass filter to attenuate harmonics. L41 and C125 form a wave trap tuned to the second harmonic. They provide additional attenuation of the signal.

SUMMARY

In this chapter you analyzed an FM transceiver, which contains a receiver and a transmitter in a single unit.

You also learned how to calculate the frequency of a crystal to receive a specific frequency.

Finally you learned how to determine the deviation required at the crystal oscillator to maintain a particular value of deviation at the transmitter's output.

SELF-TEST REVIEW

1. What is the purpose of a squelch circuit? _____

   _____

   _____

2. Refer to Schematic Diagram #3. What frequency crystal should be used at Y1 to receive a 148 MHz signal? _____

3. In Schematic Diagram #3, the output deviation is normally 5 kHz. How much deviation is required at Q9 and D3, the crystal oscillator, in order to maintain 5 kHz deviation at the output?

   _____

4. Using Schematic Diagram #3 as a reference, complete the transmitter block diagram provided in Figure 15-2. Label the frequencies present between each block, assuming a Y1 crystal frequency of 15 MHz.

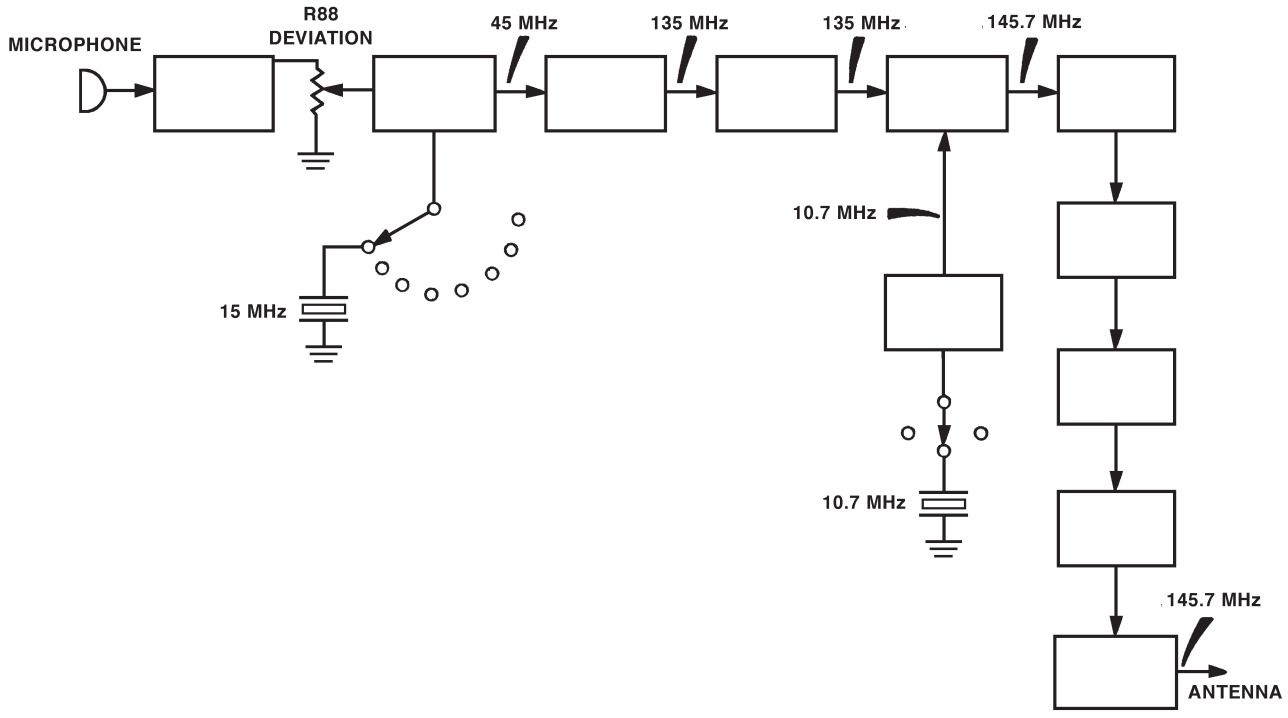


Figure 15-2

Figure for Question 4.

Section 5

Pulse Modulation

CHAPTER 16 — Analog Pulse Modulation

INTRODUCTION

As you studied earlier, analog modulation methods all involve continuously varying some characteristic of a carrier signal in step with the variations of the modulating signal. However, if you are to use pulses as the carrier, a problem is presented. Analog carriers exist at all points in time. Pulses don't. They exist only during specific periods. For the remainder of the time (the interval between pulses) pulses are effectively nonexistent.

How are you to modulate something that isn't there all the time, especially when the modulating waveform is a continuous signal that is always there? The answer lies in a technique called sampling, about which you will learn in this chapter.

Pulse modulation can be divided into two basic types: analog and digital. In analog pulse modulation, some characteristic of the pulse, its' width, height, or repetition rate, is varied in accordance with the modulating signal. In digital pulse modulation, an encoding scheme is used to send a value that represents the value of the modulating waveform at certain points in time, and not to send a modulated version of the signal at all. You will learn more about digital pulse modulation in the next chapter.

OBJECTIVES

When you complete this chapter, you will be able to:

- List the two broad categories of pulse modulation.
- Name the two types of analog pulse modulation.
- State the sampling theorem.
- Define Nyquist rate, PAM, PTM, PDM, and PPM.
- State the advantages PTM has over PAM.

SAMPLING

Sampling is just what the name implies. Rather than transmitting the entire modulating waveform, a periodic sample is taken and only the sample is transmitted. If enough samples are sent, the wave can be reconstructed at the receiving end. An example of sampling is shown in Figure 16-1. Here, a switching signal periodically energizes a relay, thereby connecting the modulating signal to the output. If the sampling signal's frequency is high enough, sufficient information appears at the output to accurately reproduce the modulating waveform. In fact, all that is needed to demodulate the output signal is a low-pass filter. This filter removes the pulses and extracts only the original information.

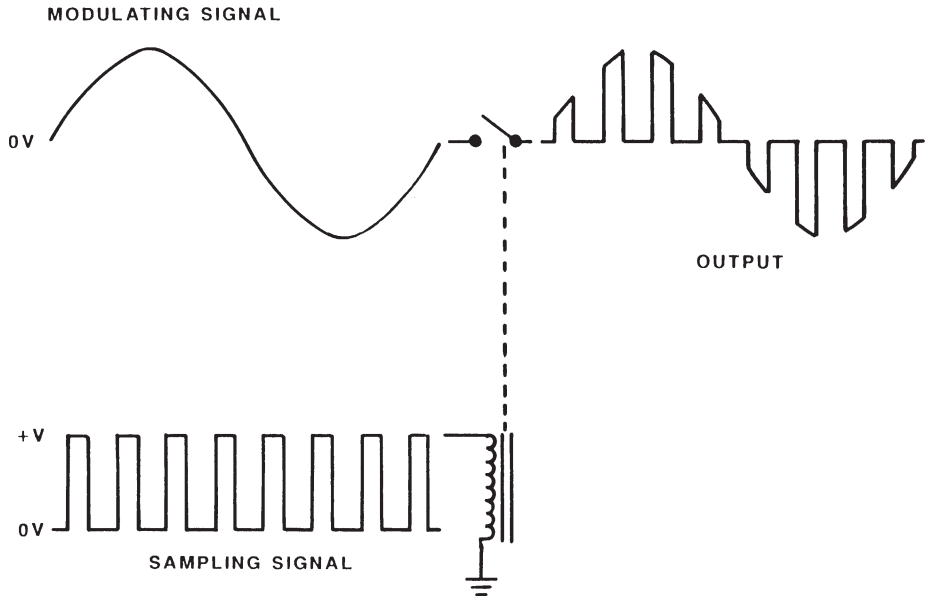


Figure 16-1
A simplified sampling system.

The critical part of this system is the sampling frequency. The minimum frequency is defined by the sampling theorem.

The sampling frequency in any pulse modulation system must be equal to, or exceed, twice the highest signal frequency in order to convey all of the information of the original signal.

The reasons behind the sampling theorem are shown in Figure 16-2. Figure 16-2A shows the bandwidth of a modulating signal with a maximum frequency of f_m. This signal is then sampled at a rate of $2f_m$ and the resulting frequency spectrum is shown in Figure 16-2B. The lower sideband extends down to f_m since it is $2f_m - f_m = f_m$. The upper sideband extends up to $3f_m$, since $2f_m + f_m = 3\ f_m$. Because the sampling signal is a pulse train, its harmonics theoretically extend to infinity, as shown by carriers at $4\ f_m$, etc. In practice, the bandwidth is much less, but still extends out far.

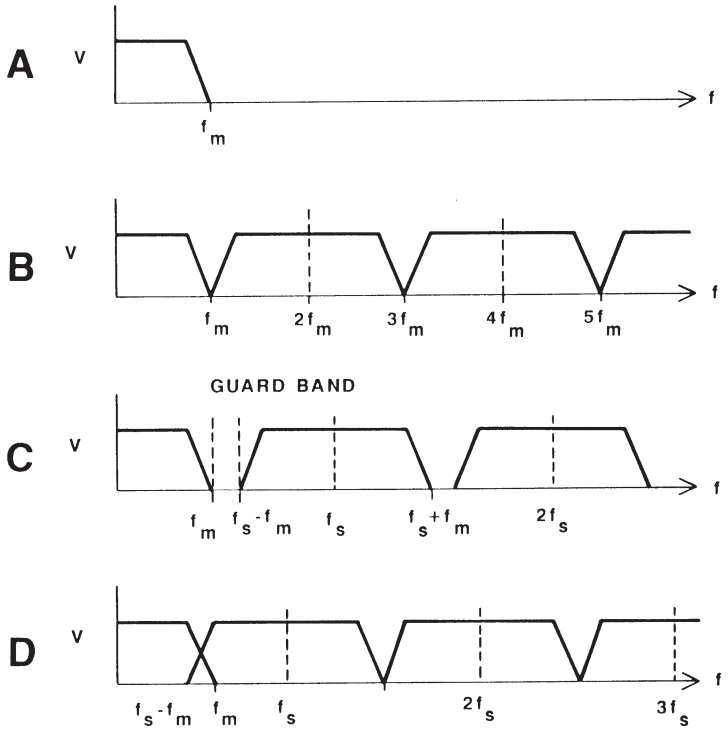


Figure 16-2

The effects of sampling frequency on the signal's frequency spectrum.

To demodulate the signal of Figure 16-2B, it is necessary to have an ideal low-pass filter with a cutoff frequency of f_m. If the filter's cutoff were infinitely sharp at f_m, it would pass the original modulating signal and nothing else.

Since ideal low-pass filters with infinitely sharp cutoffs are hard to come by, a sampling frequency of f_m is never used. Instead, a frequency slightly greater than f_m is used. The result is shown in Figure 16-2C. When the sampling frequency (f_5) is greater than f_m, a guard band appears between f_m and the lower end of the first lower sideband. This guard band allows you to use a practical low-pass filter. However, the wider the guard band, the greater the bandwidth of the pulse modulated signal.

One example of a pulse modulation system is the telephone network. Here, the voice bandwidth is limited by filters to 3.3 kHz and the sampling frequency is 8 kHz. This is well above $2f_m$ and results in a guard band of $f_5 - 2f_m$, or 8 kHz – 6.6 kHz = 1.4 kHz. This allows relatively simple low-pass filters to be used as demodulators. However, the guard band is not so wide that it unduly increases the signal bandwidth.

The situation shown in Figure 16-2D occurs when the sampling frequency is less than $2f_m$. Here, the overlap between the modulating signal and the lower sideband of f_5 means that no amount of filtering will allow an exact recovery of the full modulating signal.

Figure 16-2 proves the sampling theorem, since it shows that, in principle, the original signal can be recovered when $f_5 = 2\ f_m$ (Figure 16-2B). It also shows why the minimum sampling frequency is $2\ f_m$ (Figure 16-2D). This minimum sampling frequency is called the Nyquist rate. Figure 16-2C proves that increasing f_5 above the Nyquist rate provides a guard band which greatly eases the problems of filtering. However, increasing f_5 also increases the signal bandwidth. Therefore, a compromise must be reached that allows sufficient guard band for filtering at a minimum signal bandwidth.

All pulse modulation involves sampling. It is, therefore, essential that you understand the sampling theorem and the meaning attached to the Nyquist rate as you begin to study the details of how pulse modulation is performed.

PULSE AMPLITUDE MODULATION

The simplest form of pulse modulation is pulse amplitude modulation (PAM). In PAM, the amplitude of the pulse varies in proportion to the amplitude of the signal. This is illustrated in Figure 16-3. The modulating signal is shown in Figure 16-3A and the sampling signal in Figure 16-3B. Figure 16-3C shows a dual-polarity PAM signal. This results if the waveform in "A" is centered on a zero-volt axis or, in other words, is a true AC wave. If a PC level is added to the modulating signal, single-polarity PAM results, as shown in Figure 16-3D. In this case, sufficient DC level is added to ensure that the pulses are always positive. Likewise, a negative DC voltage could be used to obtain negative pulses.

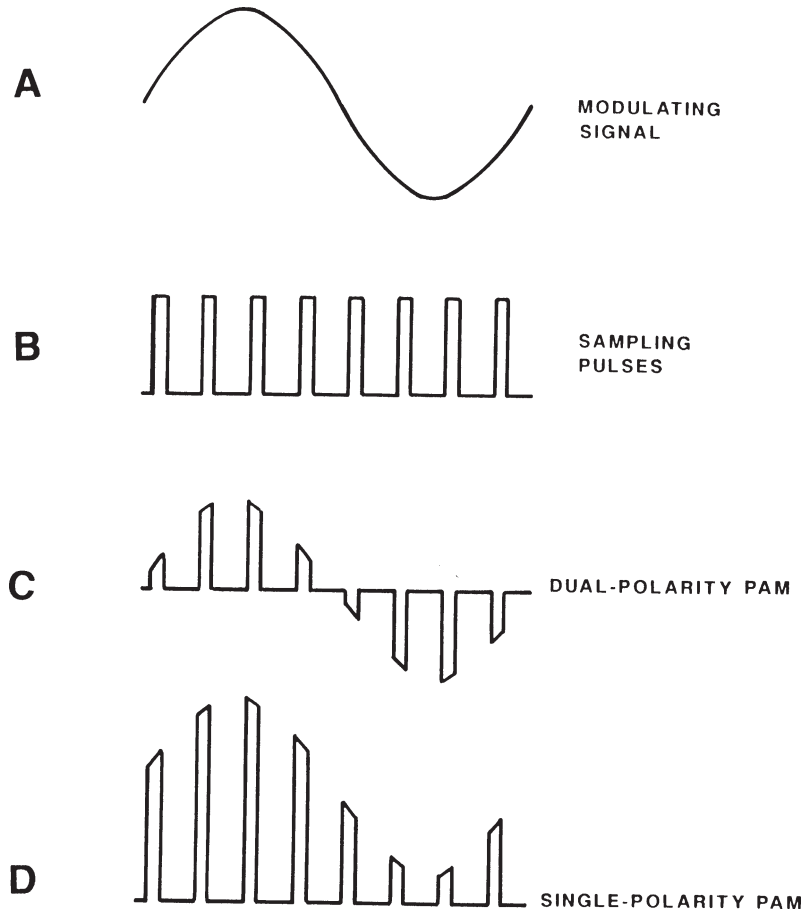


Figure 16-3
Examples of dual polarity and single polarity PAM.

One practical method of generating a PAM signal is shown in Figure 16-4. This circuit uses a 4016 integrated circuit CMOS switch. Basically, it is an FET logic switch. When the sampling pulse goes positive, the switch closes and the modulating input appears across R3 and the output. When the sampling pulse drops to zero, the switch opens and the output is zero. As shown, the circuit provides dual-polarity PAM. However, single-polarity PAM can be achieved by adding R1 and R2. These resistors form a voltage divider that adds a DC level to the input signal. The result is that the input AC wave now varies around a positive DC reference rather than a zero-volt reference.

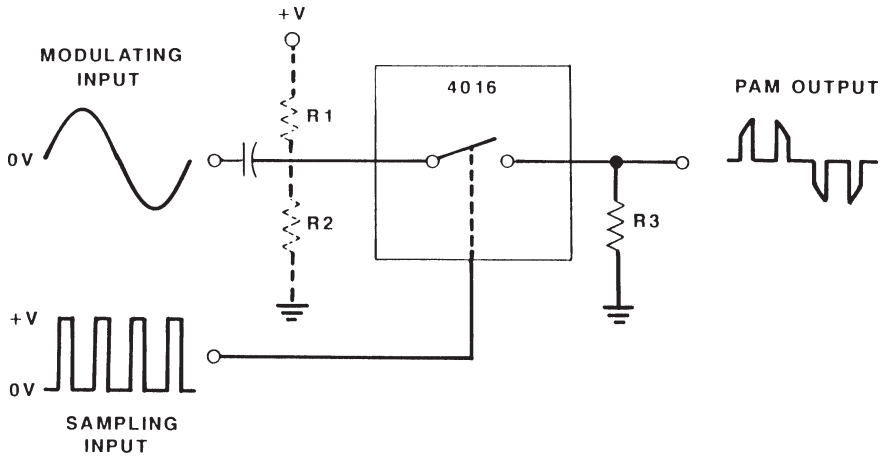


Figure 16-4
One method of obtaining PAM.

The demodulator for a PAM signal is merely a low-pass filter. It removes the sampling signal and its harmonics, and passes the original modulating signal. The rolloff of the filter must be steep enough to pass the highest modulating frequency and to fully attenuate the lowest sampling frequency component. That is, the filter's cutoff must fall well within the guard band of the particular PAM system.

When pulse amplitude modulation is used, the modulated pulses are sent over cable or wire, or are used to modulate a radio frequency carrier. When this is done, the PAM signal normally frequency modulates the carrier rather than amplitude modulates the carrier. However, PAM is not used very often to directly transmit information, since it is susceptible to noise interference.

PULSE TIME MODULATION

In pulse time modulation (PTM), the modulating signal is sampled, just as it is in PAM. However, in PTM, the amplitude of the sample is indicated by a timing variation of the modulated pulse, rather than an amplitude variation. The variable timing characteristic may be the duration, position, or frequency of the pulses. Therefore, there are three basic types of PTM: pulse duration modulation, pulse position modulation, and pulse frequency modulation. This chapter discusses the first two in detail. Pulse frequency modulation has no significant practical applications and is, therefore, omitted.

In PTM, only the timing characteristic is modulated. The amplitude of each pulse remains constant. Therefore, amplitude limiters can be used at the receiver to remove most noise pulses or spikes. For this reason, PTM has the same advantage over PAM that frequency modulation has over amplitude modulation. Thus, PTM is used much more often than PAM in communications applications.

Pulse Duration Modulation

This type of PTM is also called pulse width or pulse length modulation, however, pulse duration modulation (PDM) is the preferred term. There are three different classifications of PDM: symmetrical PDM, leading edge PDM, and trailing edge PDM. These are shown in Figure 16-5, along with the sine wave modulating signal.

Figure 16-5A shows a symmetrical PDM waveform. Here, the modulating signal is sampled and both the leading and trailing edges of the pulse are varied in accordance with the sample amplitude. When the sample is a high positive voltage, the pulse duration increases, and when it is a high negative voltage, the pulse duration decreases. When the modulating wave is at zero, the pulse is at its average or reference duration. The spacing between the center of the pulses remains constant, as shown.

Leading edge PDM is shown in Figure 16-5B. In this type of PDM, the sample amplitude varies the leading edge of the pulse. The trailing edge of each pulse occurs at the sample time and, therefore, the spacing or timing between each pulse's trailing edge is constant.

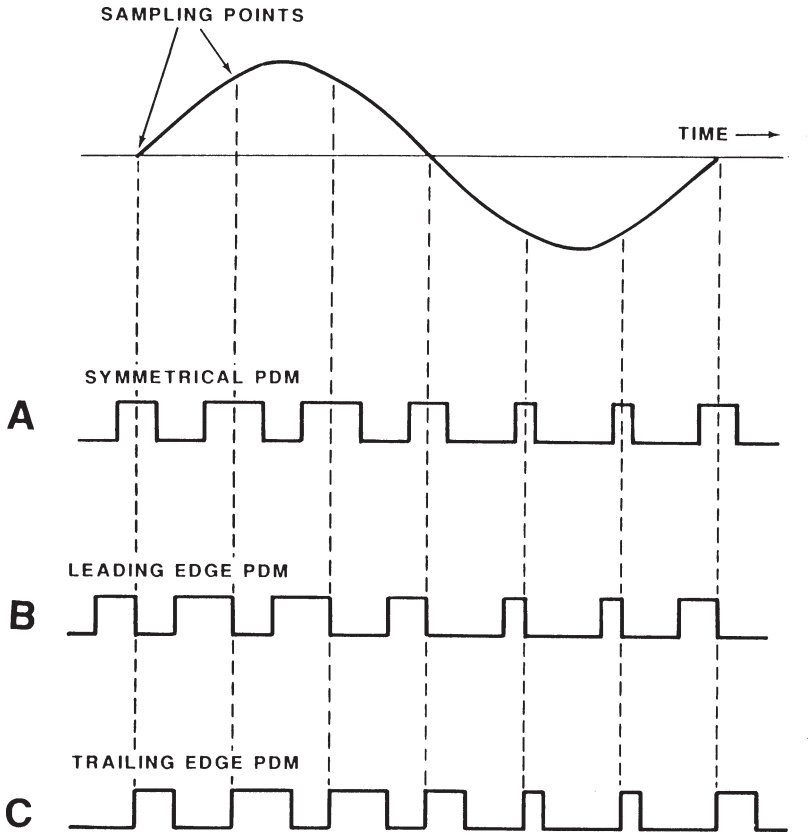


Figure 16-5

The three types of pulse duration modulation.

Figure 16-5C shows trailing edge PDM. Here, the sample amplitude varies the trailing edge of the pulse, with the leading edge of each pulse occurring at the sampling time. As with leading edge PDM, the time interval between the pulse edge corresponding to the sample time is constant.

A typical PDM generator is shown in Figure 16-6. There are two inputs: a triangular waveform at the sampling frequency and the modulating input, which is a sine wave. These two signals are summed and the output is a triangle wave superimposed on the sine wave as shown. This wave is then applied to a clipper, which removes everything below the clipping level shown. What remains is a triangle wave of varying amplitude and duration. This wave is then amplified and limited by several successive stages until all the pulses have the same amplitude, and both the leading and trailing edges are vertical. The result is a PDM wave. In this case, since the sampling wave is a triangle wave, the output is symmetrical PDM.

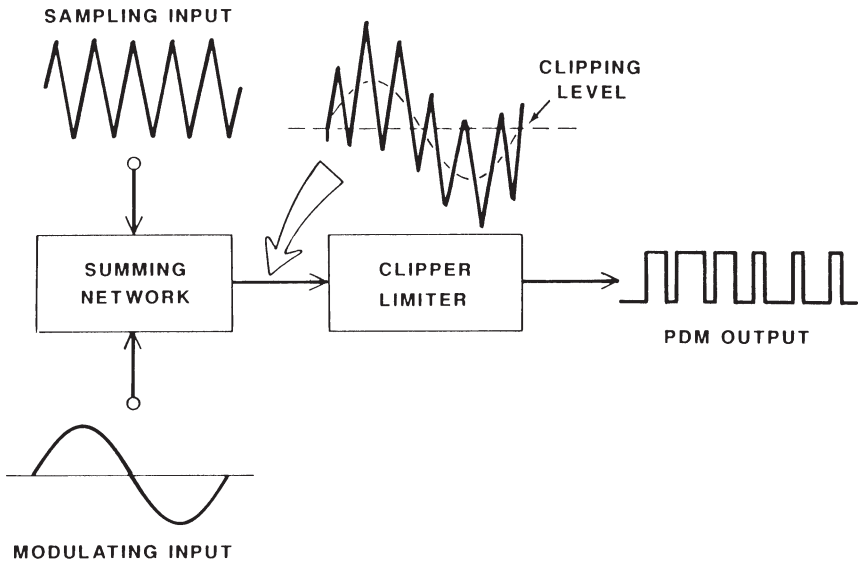


Figure 16-6

Method of generating a PDM signal.

Leading edge PDM is generated using a sawtooth sampling wave that has a sloping leading edge and a vertical trailing edge. In this way, the resulting pulse has its leading edge modulated. Likewise, trailing edge PDM is generated using a sawtooth sampling wave with a vertical leading edge and a sloping trailing edge.

Why are there three types of PDM? To detect the modulation contained in a PDM signal, it is going to be necessary to know how much the width of the pulse was varied. That means you have to be able to extract information about some point of reference that defines an unvaried pulse. In practical systems, it is often much easier to find the edge of a pulse, leading or trailing, than it is to determine where the middle of a pulse is. So, in systems where detection depends upon deriving the timing from the pulse train, leading or trailing edge PDM is easier to demodulate than symmetrical PDM, because the pulse train carries its own reference signal.

The simplest method of demodulating a PDM signal is to use a low pass filter. However, this technique can sometimes produce severe distortion. Therefore, a more elaborate but lower distortion system is normally used. One possible system is shown in Figure 16-7A. This circuit converts the PDM signal to PAM and then uses a low-pass filter to demodulate the PAM wave. This results in much lower distortion.

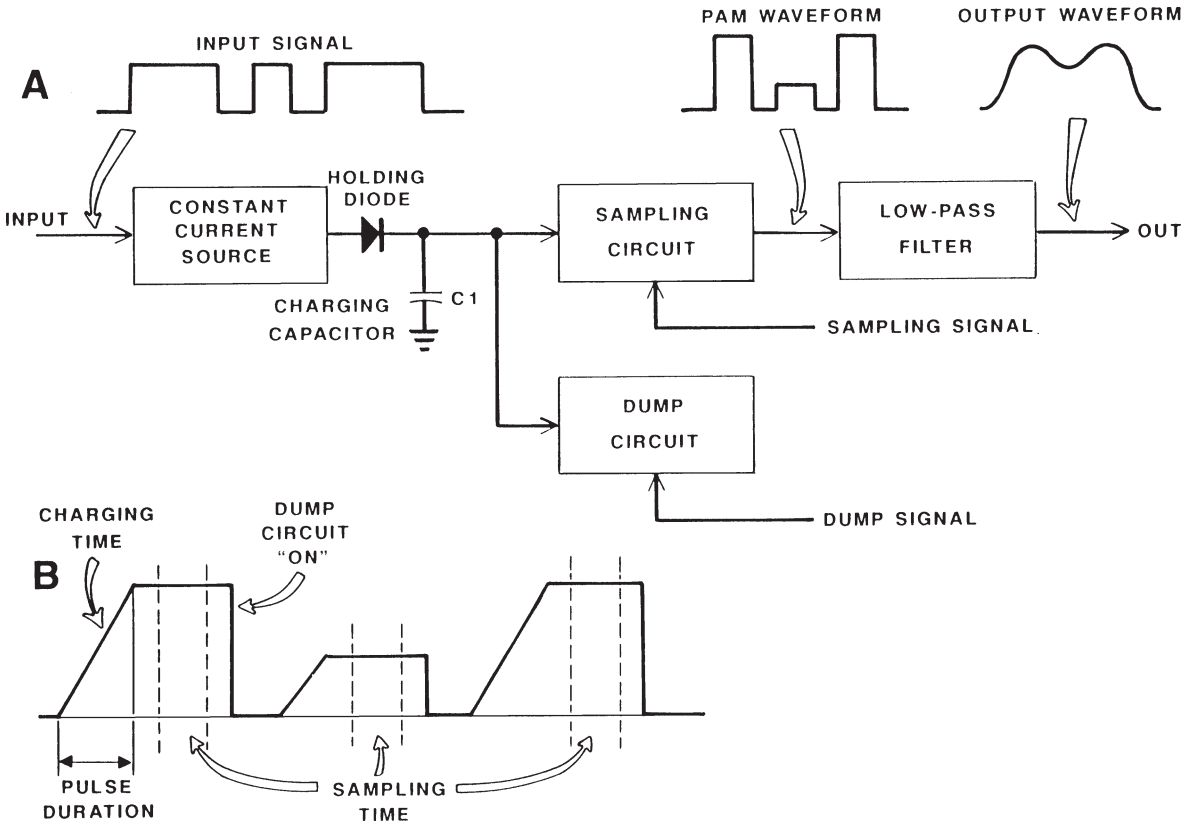


Figure 16~7

A PDM/PAM conversion demodulator.

In Figure 16-7A, the PDM input signal is applied to a constant current source. When the input goes positive, it turns on the current source, which then linearly charges C1. C1 continues to charge until the input signal drops back to zero. At this time, the current source is turned off, and the voltage is held constant across C1 by the holding diode which prevents any discharge. The sampling circuit is then turned on by a sample pulse, which is derived from the input PDM signal. As soon as the sample is taken, the dump circuit is turned on, and it discharges the capacitor, preparing it for the next pulse. The voltage waveform that appears across C1 is shown in Figure 16-7B. The output of the sampling circuit is a PAM signal, as shown. This is applied to the low-pass filter, and the original modulating signal is recovered.

Pulse Position Modulation

The next form of PTM is known as pulse position modulation (PPM). With this form of PTM, both pulse amplitude and duration remain constant, while the position of the pulse, relative to a reference pulse, is varied by the modulating signal.

Figure 16-8B shows a typical PPM waveform. The modulating signal and the reference pulses are shown in Figure 16-8A. Note that when the modulating signal goes positive, the output pulse lags the reference pulse by an amount of time proportional to the sample amplitude. Similarly, when the modulating signal goes negative, the output pulse leads the reference pulse by a proportional amount.

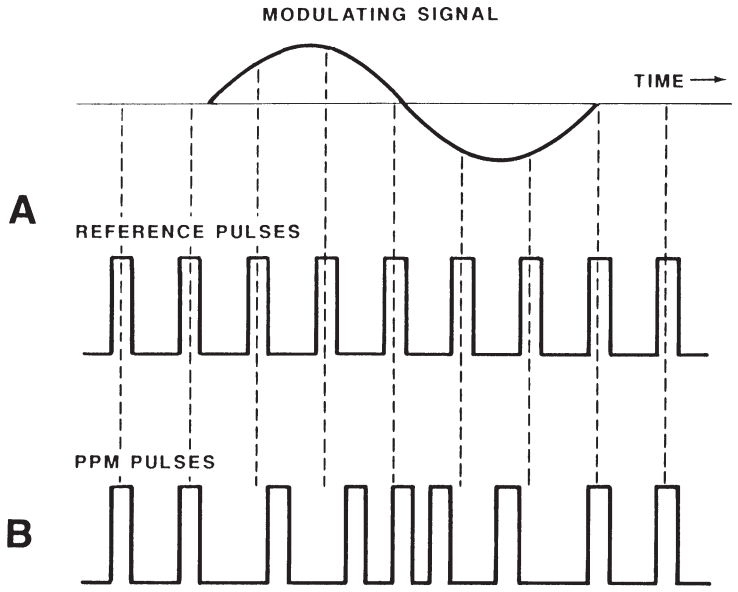


Figure 16-8

Pulse position modulation.

One method of generating a PPM signal is shown in Figure 16-9. Figure 16-10 shows the timing relationship between the various waveforms. This method uses the modulating signal to first generate a PDM waveform. In this case, it is trailing edge PDM, where the leading edge is fixed and the trailing edge is modulated. The circuit in Figure 16-9 effectively "synchronizes" a fixed duration pulse to this modulated trailing edge. It does this by first differentiating the PDM signal. The positive spikes are then clipped off, since they are fixed in time. The remaining negative spikes are synchronized to the trailing edge and are, therefore, position modulated. This clipped waveform is then

used to trigger a monostable multivibrator which generates a fixed duration pulse for each input trigger. The duration of the pulse is determined by the monostable's RC time constant.

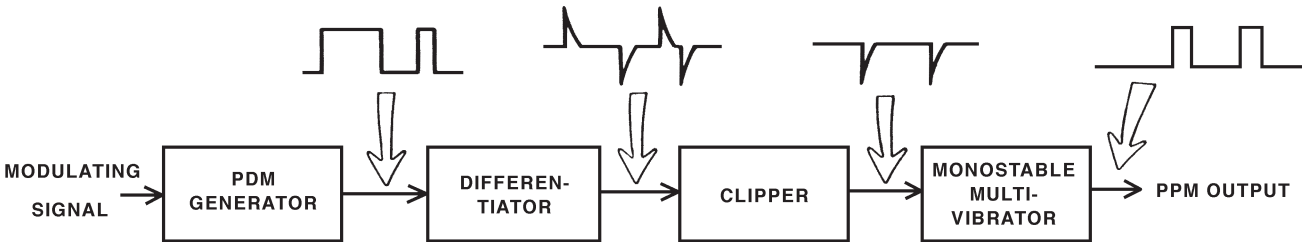


Figure 16-9

A pulse position modulation circuit.

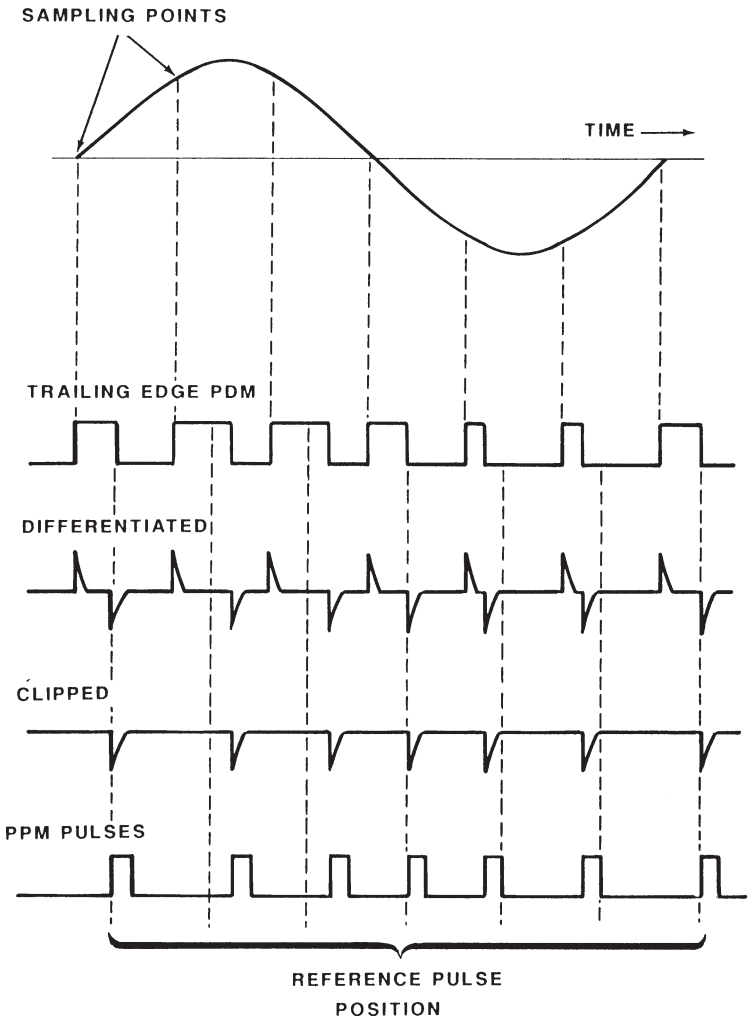


Figure 16-10

Waveforms for the PPM generator.

Figure 16-11 shows a typical PPM demodulator, and Figure 16-12 shows the timing relationship between various signals. The PPM signal is first applied to a limiter stage to remove any noise, and to effectively regenerate the signal. It is then sent to a circuit that removes the synchronizing pulses sent by the transmitter. These pulses are used to lock the local reference pulse generator to the correct phase and frequency to demodulate the PPM signal. The reference pulses and the PPM signal are then applied to a flip-flop. Note that the incoming PPM pulse lags the reference pulse by an amount proportional to its position modulation. As shown, in Figures 16-11 and 16-12, the reference pulse turns on the flip-flop and the PPM pulse turns it off. Therefore, the output of the flip-flop is a PDM wave whose pulse duration is directly proportional to the position of the input pulse. The resulting PDM signal is then converted to PAM as described earlier. The PAM wave is then coupled to a low-pass filter to recover the original modulating signal.

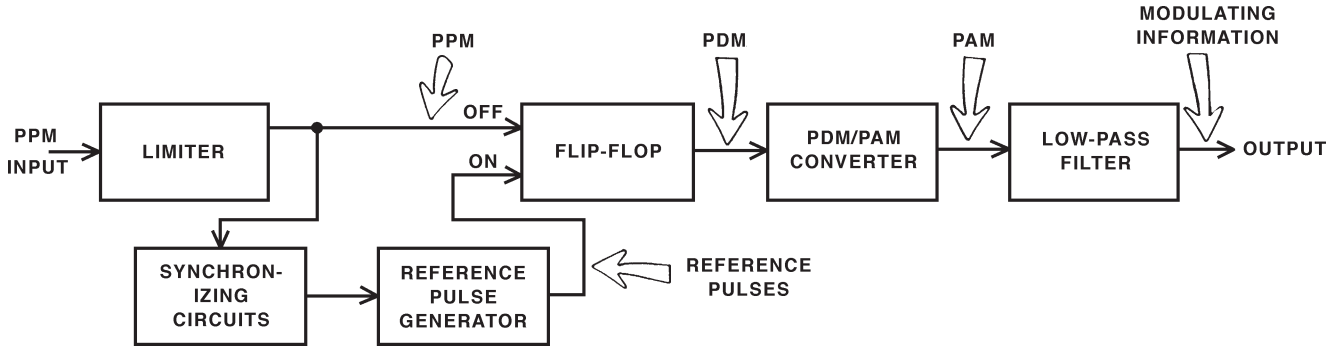


Figure 16-11

One method of demodulating a PPM signal.

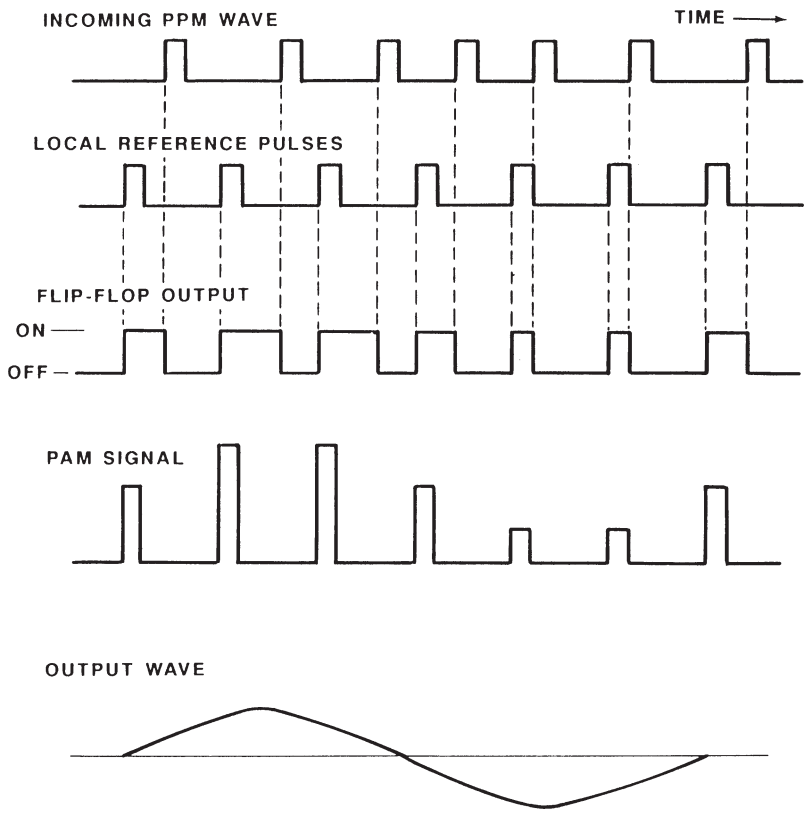


Figure 16-12

Waveforms for the PPM demodulator.

PPM is widely used because it is relatively immune to noise interference on the transmission channel, and because detection is straightforward and can readily regenerate a "clean" pulse train. To correctly receive PPM, it is necessary only to detect the presence or absence of a pulse. Amplitude is unimportant, so limited-bandwidth and noisy channels have little effect on the fidelity of the demodulated information using PPM.

SUMMARY

In this chapter you studied analog pulse modulation. You learned that analog pulse modulation is very much like AM and FM as studied for analog signals.

Sampling is used to break up a modulating waveform into pieces that are then transmitted. As the number of samples increases, the easier it is to reconstruct the original waveform.

Pulse amplitude modulation, where the amplitude of the pulse varies in proportion to the amplitude of the signal, is the simplest form of pulse modulation. One disadvantage of pulse amplitude modulation is its susceptibility to noise.

In pulse time modulation, the modulating signal is also sampled, but the amplitude of the signal is indicated by a timing variation of the modulated pulse. Since the amplitude of each pulse remains constant, limiters can be used at the receiver to reduce noise pulses or spikes.

Pulse duration modulation was discussed next. With symmetrical PDM, the both the leading and trailing edges of the pulse are varied by the sample amplitude. In leading edge PDM, the leading edge of the pulse varies with the sample amplitude while the timing between each pulse's trailing edge remains constant. Conversely in trailing edge PDM, the trailing edge of the pulse varies with the amplitude of the sample amplitude while the timing between each pulse's leading edge remains constant.

Finally pulse position modulation was discussed. With this method, the pulse amplitude and duration remain constant while its position is varied by the modulating signal.

SELF-TEST REVIEW

1. What are the two broad categories of pulse modulation? _____

   _____

2. What are the two basic types of analog pulse modulation?

   _____

3. What is the sampling theorem? _____

   _____

   _____

   _____

4. What is the Nyquist rate for a modulation signal whose highest frequency is 8 kHz? _____

   _____

   _____

   _____

5. What is PAM? What are the two different types of PAM?

   _____

   _____

6. A _____ _____ filter can be used to demodulate a PAM signal.

7. What is PTM? What are the three basic types of PTM?

   _____

   _____

   _____

8. What advantage does PTM have over PAM? _____

   _____

   _____

9. What is PDM? List the three types of PDM. _____

_____

_____

_____

10. What is PPM? _____

_____

CHAPTER 17 — Digital Pulse Modulation

INTRODUCTION

In the previous chapter you learned about analog pulse modulation. Analog pulse modulation schemes like PAM suffer from the same problems as their analog cousins: they are susceptible to noise and distortion during transmission.

You should know from your previous study of digital electronics that if you can detect that a pulse is present, you can regenerate it to create a "clean" pulse. If you don't need to know exactly how wide the pulse originally was, or what its amplitude was, you can also get rid of most of the noise that was picked up along the transmission path. In fact, unless a burst of noise caused you to inaccurately interpret the existence of the pulse, you can get rid of all of the noise that has accumulated. The trouble is, you need to know precisely the pulse characteristics that the noise affects to demodulate analog pulse modulation.

Wouldn't it be nice if you could devise a digital signal that told everything that was needed to reconstruct the original waveform? Then you could send the signal by any of the modulation techniques that seemed appropriate, and capitalize on our ability to regenerate the pulses at the receiver. Since the pulses wouldn't actually be the signal, but merely the instructions on how to recreate the signal, you could rebuild the original signal with practically no distortion or noise at all.

Well, such a scheme exists. It is called digital pulse modulation. The most common form of digital pulse modulation (DPM) is pulse code modulation, or PCM, for short. PCM is not so much a modulation scheme as it is a coding system. Nevertheless, it is called modulation, and everyone treats it that way, so you shall, too. This section introduces you to pulse code modulation: what it is, how it is generated, and how it is detected. But, before you can understand PCM, you must first learn about quantization and how it works.

With PCM, the modulating signal is sampled and then quantized. In quantization, each sample is assigned a specific numerical value according to its amplitude. This numerical value is then represented by a group of pulses of equal amplitude and duration. The absence or presence of pulses represent the modulating signal's value in the binary number system. This system has many advantages and, therefore, has many applications in modern communications.

OBJECTIVES

When you complete this chapter, you will be able to:

- Define digital pulse modulation, quantization, and quantization noise.
- List the three basic types of digital pulse modulation and describe each one.
- Describe how PCM is generated and detected.
- Explain the benefits of DPM.
- Compare and contrast analog and digital pulse modulation.

QUANTIZATION

The key to PCM is the concept of quantization, which simply means dividing something up into parts. In our case, you are going to divide things so that there are only so many specific values that the signal can have. To illustrate this concept, let's consider the case of pulse amplitude modulation shown in Figure 17-1.

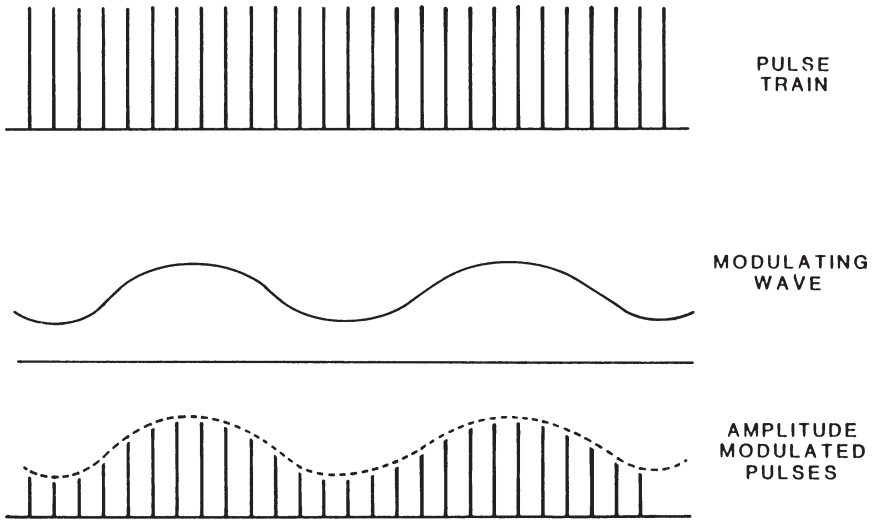


Figure 17-1

Pulse amplitude modulation by a sine wave.

You can see from the resulting string of modulated pulses that the height of each pulse is proportional to the amplitude of the modulating signal. That, after all, is the very basis of PAM. What is not quite so obvious, is that the pulses can have an infinite number of heights. Since the input signal can vary in any manner, and since each pulse is proportional to the value of the input at that moment, there is no limit to the number of values that the modulated pulses can have. Clearly, that can pose some problems. Our ability to detect very slight differences in pulse amplitudes is limited. Again, you run into the problem of infinity. If our detector can detect changes in level of a millivolt, then if the change from one pulse to another is one-tenth of a millivolt, it cannot be detected. If you cannot detect all possible pulse amplitudes, there is clearly a limit to the fidelity of the PAM system.

What to do? We can never devise a digital detector that can handle all possible levels of pulses.

Let's try another approach. Suppose you deliberately limit the values that the pulses can have. Consider Figure 17-2. This is the same situation as shown in Figure 17-1, except, you have arbitrarily decided that the output pulses can have only one of four possible values. We do that by requiring that an input signal between the values shown produce a pulse of a certain amplitude. That pulse is proportional to the signal level, half-way between the limits you set for the input. Looking at the output pulse stream, you see that it resembles the shape of the modulating signal, but not very closely.

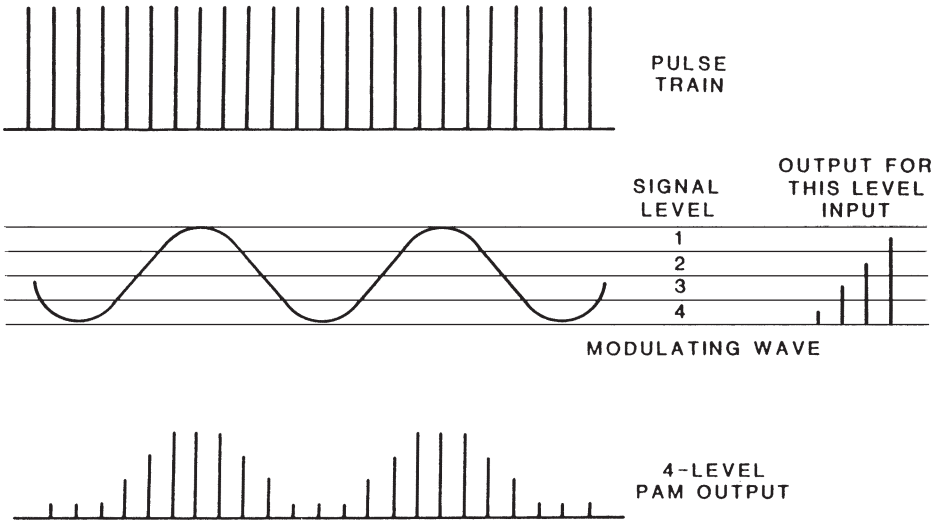


Figure 17-2

4-level PAM.

Suppose now that you increase the number of levels that the pulses can have, to eight. That means, that you need to be able to decode eight amplitude levels on the input signal. Figure 17-3 shows this approach. Notice that the output pulse envelope resembles the input modulation much more closely than was so for the 4-level scheme shown in Figure 17-2.

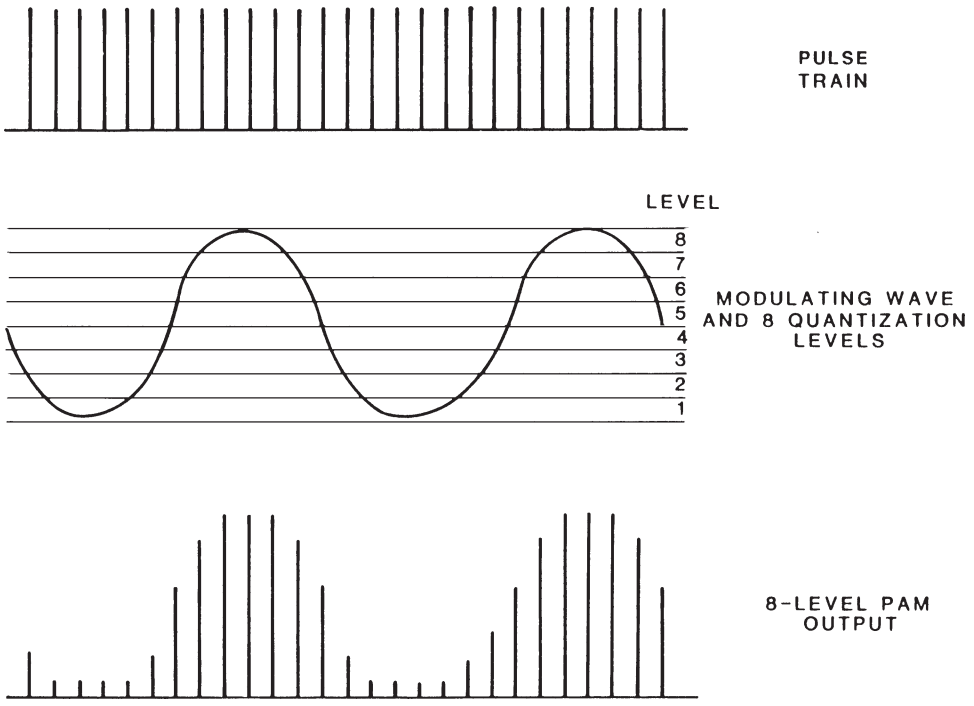


Figure 17-3

8-level PAM.

You could increase the number of possible levels, and you would observe that as the number of levels increases, the representation of the modulating signal improves. That should not be surprising, since you started out with an infinite number of levels, and those gave a good representation indeed.

What you have done is quantize the input signal by saying that any value of the input that lies between two limits will be represented by a pulse of a predetermined height, that is related to the value of the input signal. What you are effectively doing is setting-up a series of amplitude bands, or slices, that represent different input levels. For example, the pulse amplitude might be equal to the average value of

the input over that range. As an example, if one quantization band is from 1 to 2 volts, any input value in that band would be represented by a pulse height proportional to an input of 1.5 volts.

Obviously, by quantizing the signal you have decreased the resolution possible in detecting it. Put another way, you have limited the fidelity with which you can reconstruct the signal at the receiver. This does not seem like progress. Why would you want to do such a thing?

Well, if there are only so many values that the modulated pulses can have, you know from the outset what the possible values are. If only pulses with amplitudes at tenth-volt multiples are possible, and you detect a pulse of 1.02 volts you can safely assume that it should really have been 1.00 volts, and correct accordingly. In this way, you can eliminate some noise and distortion.

Getting rid of noise sounds good, but is it worth the decreased fidelity caused by quantization? It can be. Remember our discussion about representing square waves by several sine waves. You found that after a point, no practical improvement was made by adding more sine waves. The shape of the resulting square wave improved, but not the ability to detect it. Much the same thing is true with quantization. A trade-off must be reached between the number of levels that provides adequate fidelity of received signal, and the number that provides a realistic number of discrete values to correct for transmission noise. It turns out that such trade-offs are practical, enabling a capability to take advantage of the benefits you have discussed without incurring an intolerable loss of signal fidelity. The telephone system does this regularly. Although the human voice has frequency components well outside the 300-3400 Hz frequency band, this portion of the voice signal is adequate for the fidelity required of the telephone, and is all of the voice signal that is transmitted.

QUANTIZATION NOISE

After the input signal has been quantized, it is impossible to reconstruct it **exactly**. The deviation of the reconstructed signal from the original signal is called quantization noise. The amplitude of this noise can never exceed one quantization level, and is therefore easier to deal with than randomly occurring natural noise, which can be very strong indeed. The more quantization levels there are, the lower the maximum possible value of the quantization noise.

When PCM is used to transmit voice, quantization noise can sometimes be heard as a low-pitched buzz superimposed on the voice signal. It is a particular problem when it represents a significant fraction of the signal amplitude, as on the weaker input signals. Because of this, systems have been created to minimize quantization noise effects.

VARIABLE-LEVEL QUANTIZATION

Consider the situation of Figure 17-4. Using the same quantization levels as were used in Figure 17-3, the weak signal shown will not be quantized in such a way as to permit its accurate reconstruction. This is because of the distance between the quantizing levels, not the fact of the input signal being weak. Think of it this way: with a weak signal and fixed quantizing level intervals, the input signal does not cross very many quantization level boundaries, and therefore relatively few different levels are observed in the modulated pulses. You could simply amplify the signal, but even then the quantization would be in error; while the shape of the signal would be more accurately represented, its amplitude would be grossly distorted.

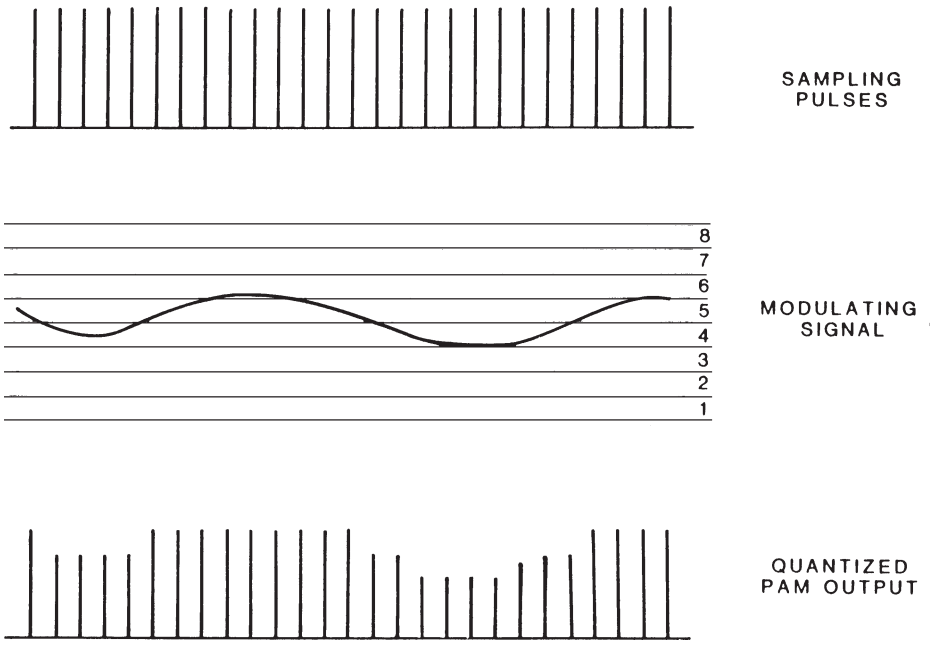


Figure 17-4
Quantizing a weak signal.

Nothing demands that all of the quantizing levels be the same width, though. If you were to make the quantization slices smaller where the signal is weaker, and larger where it is stronger, you could improve the representation of the signal in the pulse train. You can see this by looking at Figure 17-5.

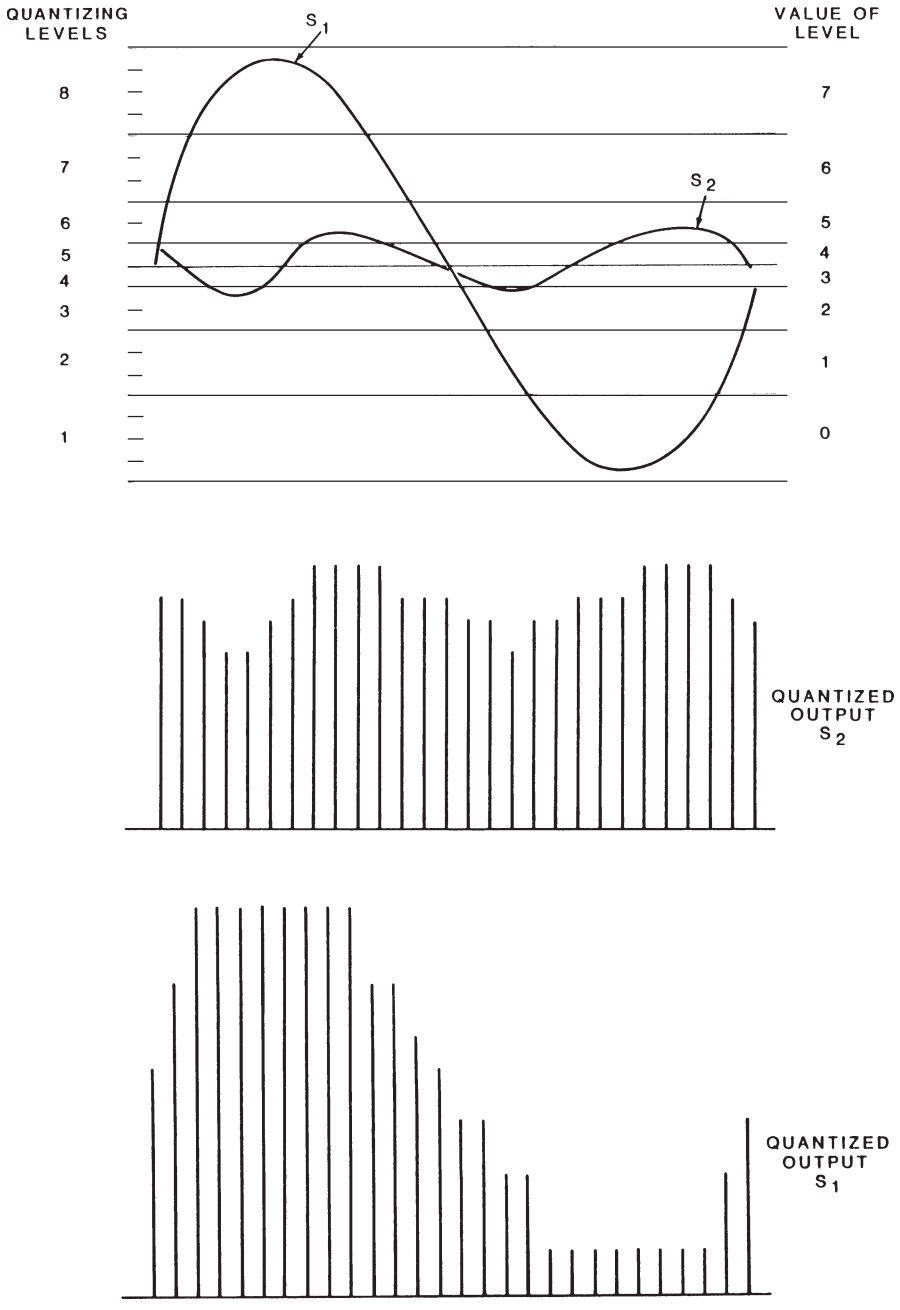


Figure 17-5

Variable-level quantization.

By choosing smaller quantizing intervals for the weaker signal ranges, you force the signal, even a weak one, to cross more quantization levels than it would with fixed spacing of levels. This, in turn, provides more values of the modulated pulses that describe the signal. The result, as you see, is that the representation of the weak signal is more accurate than in the case of fixed spacing between the quantizing levels. Notice also that the representation of the strong signal does not suffer appreciably from the use of varying spaces between the quantizing levels.

This procedure of changing the quantizing levels is called companding, from a contraction of COMpression and exPANsion. The physical devices that do the companding are called compandors. Standards have been developed to provide guidance on companding levels and procedures. Two primary standard specifications, called A-law and mu-law companding, have been developed for companding used in PCM transmission, and these are widely used. More about this later.

MORE ABOUT SAMPLING

In quantizing the input signal, you have actually sampled it. We have taken a sample of the value of the input signal at discrete intervals in time- when each of the pulses occurs—and have used that value to modulate the pulse that occurs at that time.

The most common application of PCM is in the transmission of speech. How many samples and quantizing levels are needed to do this? Typically, speech signals in telephone networks are limited to a maximum frequency of not more than 4 kHz. Sampled at the Nyquist rate, a 4 kHz signal requires 8,000 pulses per second to properly sample it. It has also been shown that the ability to distinguish 128 volume levels provides an adequate ability to reconstruct speech. So you need a pulse train of 8,000 pulses per second (pps) and 128 quantizing levels to transmit speech using PCM.

CODING

So far, you have examined quantizing levels and sampling rates, but you haven't really defined PCM. How do these two concepts fit together to make a modulation or coding scheme?

Coding is the critical phase. In PCM, you do not transmit the actual modulated pulses. Remember that each quantized pulse can assume only one of the valid levels determined by our choice of quantizing levels. A code is generated for each of the amplitude modulated pulses according to its value. If it represents quantization level 6, the binary code for 6 is encoded and transmitted, not the modulated pulse itself. Figure 17-6 shows how this applies in our simplified example with eight quantizing levels.

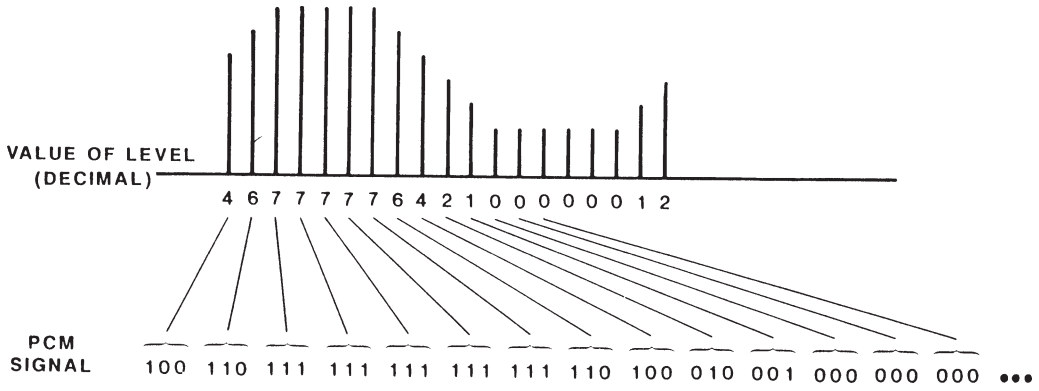


Figure 17-6

PCM with 8 quantizing levels.

This, then, is PCM. The modulating signal is sampled and quantized, and a digital code that tells the value of the modulating signal at each sampling point is transmitted, not the modulated pulse itself. The amplitude and frequency of the transmitted pulses have nothing to do with the amplitude and frequency of the modulating signal, they only represent its value. Therefore, you need only to detect the existence of the transmitted pulses at the receiver to recover the coded values. Given the coded values, you can reconstruct the input signal with exactly the same fidelity that was possible from the original stream of PAM pulses immediately after modulation.

The number-of-pulses per second needed to adequately sample the input waveform is that required for the PAM phase, not for transmission. Each PAM pulse must be represented by a transmitted code that gives its value. The number of bits needed to do that depends upon the number of sampling levels. For our simplified 8-level example, 3 bits are needed to represent the value of each (it takes 3 bits to count up to 8 in binary). For the case of 128 levels that you examined earlier for telephone usage, 7 bits are required ($2^7 = 128$). So, to transmit a 4 kHz telephone speech signal, you must send 7 bits for each of the 8,000 pulses per-second-needed to sample the original signal, or $7 \times 8000 = 56{,}000$ bits-per-second. This is the rate seen by the transmission channel. The channel does not see merely the sampling rate. The greater the number of quantization levels, and the higher the frequency of the input signal, the greater is the pulse rate required to transmit it using PCM.

PCM TRANSMITTERS AND RECEIVERS

A simplified diagram of a PCM transmitter is shown in Figure 17-7. It consists of two sections. The first section is an analog-to-PAM converter that performs the quantizing process. The second section is a PAM-to-PCM converter that does the binary coding. The total system is actually an analog-to-digital converter.

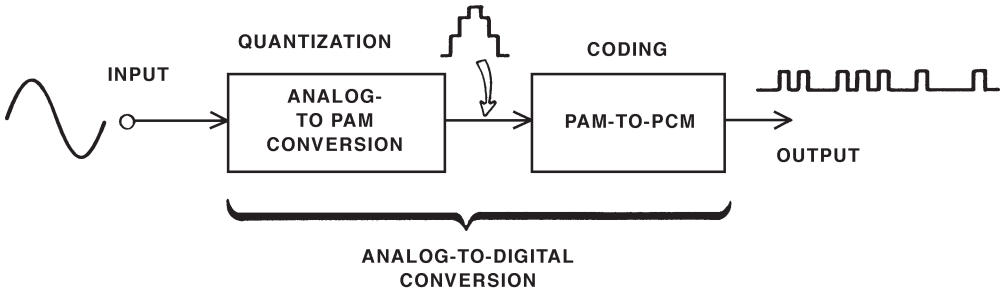


Figure 17-7

A PCM receiver is shown in Figure 17-8. It is made up of a PCM-to PAM converter and a low-pass filter to convert that PAM back to the original modulating signal. It is, in essence, a digital-to-analog converter.

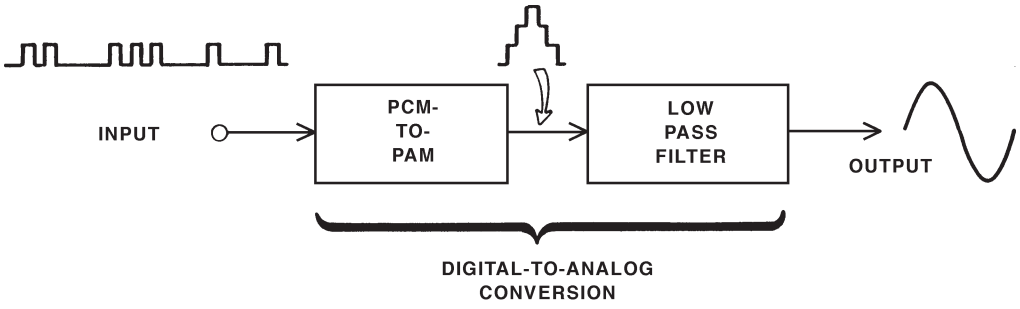


Figure 17-8
A PCM receiver.

As mentioned earlier, the primary advantage of PCM is its much better immunity to noise and interference. For example, a typical PCM transmission can be sent over a communications channel having a signal-to-noise ratio of 21 dB with minimal error. In fact, the error would be just one pulse missed or decoded improperly every 17 minutes. If the signal-to-noise ratio is improved to 23 dB, the error rate drops to one error every four months! To achieve this low error rate in an AM system would require a signal to noise ratio of 60 to 70 dB.

OTHER DPM SYSTEMS

There are other digital pulse modulation systems that are variations of PCM, or slightly different systems. Two of the more common types are differential PCM and delta modulation.

Differential PCM is very similar to ordinary PCM. However, in this system, each binary word indicates the difference in amplitude between the present sample and the previous one. It specifies whether the change is positive or negative and the magnitude of the change. Therefore, the PCM code needs to indicate only the relative amplitude of the sample rather than the absolute amplitude as in normal PCM. This system takes advantage of the redundancy of most modulating signals, such as speech. That is, each sample is related to the previous sample; large variations from one to the next are unlikely. For this reason, it takes fewer bits to indicate the relative change in amplitude

than the absolute amplitude. Thus, differential PCM has a smaller bandwidth than normal PCM. However, complications in encoding and decoding can outweigh any bandwidth advantage.

Delta modulation is basically a 1-bit differential PCM system. In the simplest form of delta modulation, just one bit is sent per sample and its polarity indicates whether the signal is larger or smaller than the previous sample. An example of this is shown in Figure 17-9. Figure 17-9A shows the quantized wave and Figure 17-9B shows the delta modulation pulse train.

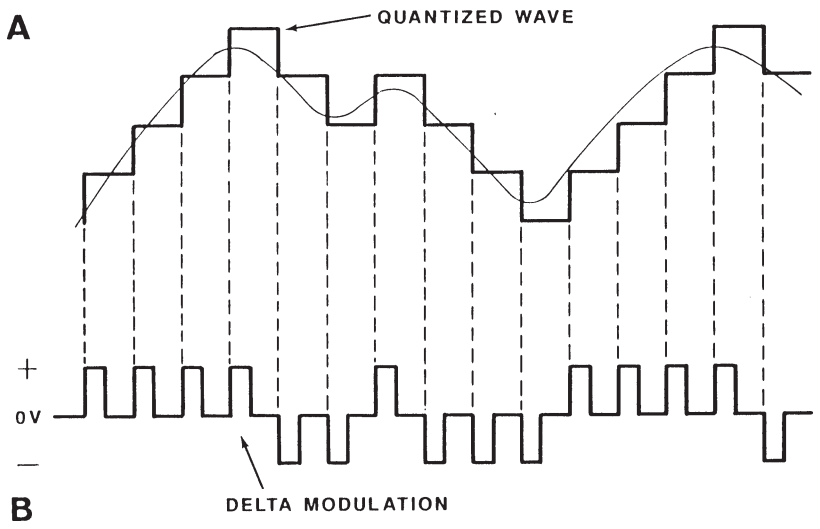


Figure 17-9

Delta modulation.

This system has several advantages: greatly simplified encoding, decoding, and quantization. However, it cannot handle rapid amplitude variations between samples unless the sampling frequency is very high. Therefore, delta modulation requires a higher sampling rate and larger bandwidth than a comparable fidelity PCM system. Its advantages somewhat offset this disadvantage and, as a result, its use is increasing.

ANALOG AND DIGITAL PULSE MODULATION COMPARED

Analog pulse modulation suffers from all the disadvantages of its analog counterparts, AM and FM. Noise and distortion on the transmission path can seriously affect analog pulse modulation. PAM is most affected by noise, which affects the amplitude of the pulse just as it affects analog AM signals. Once the noise has been added to the pulse, it is impossible to remove it or to regenerate the pulse without the noise.

PTM is more resistant to noise. Limiters can be used to remove the noise that affects the pulse amplitude. However, PTM requires that the leading and trailing edges of the pulses be accurately determined to demodulate the signal. Noise and distortion combine to distort the location of the pulse edges and to mimic pulses by noise bursts, all of which result in noise and distortion in the received signal.

For these reasons, analog pulse modulation is generally not used on long transmission paths. It is invaluable, however, as the initial step in obtaining digital pulse modulation (DPM). In practice, most DPM is actually one form or another of PCM.

PCM is becoming very widely used, especially in the telephone industry. Virtually all new telephone equipment placed in service within the past 20 years (except for that maintaining compatibility with older equipment) has been PCM equipment. So rapid has been the transition that it is referred to as the "PCM Revolution."

Clearly, for PCM to be so popular, it must have some significant advantage over competing schemes. You saw that generating a PCM signal is more complex than any of the other modulation schemes we've described. In fact, you must go through a PAM process to even begin creating a PCM signal. What are the advantages of PCM, and why is it becoming widely used in place of AM, FM, and PM?

The most significant advantage of PCM is that by using it, you are able to limit the noise component of the signal to the value of the quantizing noise. With analog signals (such as AM, FM, and PM), noise is additive. Each time the signal is amplified, the noise that has been added along the way is amplified together with the desired signal, and becomes an inseparable part of it. With PCM, noise does not add to the signal. At each repeater, the signal is completely regenerated, which removes any noise that had been a part of the signal before that point.

At the last receiver, the original signal is reconstructed by using the code sent in the PCM signal, not from the pulses themselves. This reconstruction can be just as accurate as the system design (quantizing levels and sampling rate) allows, and is not reduced in fidelity by noise on the transmission channel. That is a significant advantage, and is the one that began the "PCM revolution" in which we find ourselves.

Another advantage is cost. The great advantages in large-scale integration (LSI) of digital integrated circuits (IC) have made it cost effective to process digital signals. It is now cheaper to build digital circuits than it is to build analog ones, and as PCM primarily involves digital technology, this economy can be used to advantage. PCM modulators can be built on a chip if necessary, and the pulse regenerators are straightforward digital devices which are inexpensive to make in quantity. Repeaters are needed at closer intervals for PCM than for analog transmission, but the decreased cost of the repeaters still makes the overall system cheaper than the competing approach. So the second major advantage of PCM is that it is less expensive than most competing approaches.

More and more of the world's communication is being done with computers. Computers communicate digitally. PCM systems are built to transmit and receive digital pulses accurately. It should be possible to put computer pulses directly onto the PCM transmission media, and obtain the same benefits of pulse regeneration and low-cost transmission as the PCM source signal enjoys. This can be done, making it easier and more efficient to carry computer communications on a digital PCM system than on an analog system. Thus you have uncovered a third advantage of PCM: it facilitates digital communications at precisely the time that those communications are increasing in importance.

Privacy is a sizable problem in communications. An increasing number of communications users require the capability to protect their transmissions from eavesdropping by casual listeners. Governments have always been interested in secret communications for certain applications, but today even businesses and private parties are interested. Without secure communications, computer-to-computer communications can be at the mercy of anyone able to drop in on the line and mimic one of the computers. Simple applications like automatic bank teller machines, which communicate with a central computer, offer major opportunities for illicit activities unless their communications are secure.

Encryption systems are now available to the general public. They are invariably digital devices. If your signal is digital, it can be enciphered readily; if you want to encipher an analog signal, your problems have only just begun. PCM, being digital, is easily encrypted. This is an increasingly important advantage.

These, then, are the four major advantages of PCM over competing transmission schemes:

1. Performance
2. Cost-Effectiveness
3. Digital Compatibility
4. Easy Encryption.

Does this mean that all other systems will disappear? Not at all. As you shall see, a signal may be modulated and demodulated many times, using many different techniques, on its route from sender to receiver. There is a tremendous investment represented by the installed FDM and TDM systems that have been put in place over the past half century. The magnitude of this investment guarantees that these older systems will be around for a long time. Nevertheless, PCM is an extremely important modulation and coding technique. While it is not a panacea, and it will not replace all other systems, it is probably the most significant development in modulation schemes in recent times.

The forms of analog pulse modulation you have studied are most often used to generate PCM or another type of DPM. The advantages described above are of such significance that DPM is virtually the only kind of modulation being used in new, non-broadcast communications systems.

SUMMARY

In this chapter you learned about digital pulse modulation, which is actually a coding scheme. Analog pulse modulation is used to quantize an input signal. Then a code representing the value of the signal is transmitted and decoded, rather than sending a sample of the amplitude of the signal itself, as is done in AM and FM systems.

SELF-TEST REVIEW

1. What is digital pulse modulation? _____

   _____

   _____

2. Define quantization. _____

   _____

   _____

3. What is quantizing noise and how can it be reduced? _____

   _____

   _____

4. List the three basic types of digital pulse modulation and describe one. _____

   _____

   _____

   _____

   _____

   _____

   _____

   _____

   _____

   _____

5. PCM

   a. Begins with PAM.

   b. Is a coding system.

   c. Does not transmit the actual signal that is the input.

   d. All of the above.

6. Modern digital technology is largely responsible for the attractiveness of PCM in practice. TRUE or FALSE? _____

7. PCM can get rid of noise that has built up on the transmitted multiplex signal as it travels between two locations. TRUE or FALSE? _____

8. By impressing a noisy analog signal onto a PCM system, you can eliminate the noise that has so far been incorporated into the signal. TRUE or FALSE? _____

9. The easiest type of system to use for secure (encrypted) computer communications is:

   a. FDM

   b. PCM

   c. PDM

   d. PPM

CHAPTER 18 — Multiplexing

INTRODUCTION

Until now, you have been examining communications that involve one transmission path for every signal. While that approach is workable, it is also inefficient and costly. Using a process known as multiplexing, you can put many messages on the same channel without having them interfere with one another. This provides for much greater economies in information transmission than the "one signal, one channel" approach. In this chapter, you will examine the three primary approaches to multiplexing and how they affect data communications.

OBJECTIVES

When you complete this chapter, you will be able to:

- Define multiplexing.
- Name the basic types of multiplexing and define each one.
- Define frame, commutator, and decommutator as they apply to multiplexing.

THE 4-kHz CHANNEL: A BASIC BUILDING BLOCK

In the early days of telephony, when each conversation was carried by its own pair of wires, not much thought was given to bandwidth. More than enough was available, as the bandwidth of a pair of wires is quite high. When multiplexing began, however, it became necessary to find out how much space each conversation required. If you are going to fill a box with bricks, you need to know how large each brick is if you are going to predict how many the box will hold, don't you?

Considerable analysis of the human voice was conducted, since the primary problem was sending conversations over telephone circuits. This analysis found that the average human voice contains frequencies from about 100-8000 Hertz, with women's voices having a higher average frequency than mens'. It also found that most of the power in the human voice was contained in the band from 300-2800 Hertz, and that if this band of frequencies were transmitted, the voice was both recognizable (that is, you could tell who the speaker was) and intelligible (meaning that you could understand what was said). It is plainly more attractive to transmit a 3 kHz wide signal than it is an 8 kHz signal, if bandwidth is limited, so the telephone industry immediately took steps to limit the bandwidth of the signal impressed onto the network to about 3 kHz. For obvious reasons, each voice circuit, or channel, came to be known as a "3 kHz channel."

Over the years, it was found advantageous to widen the bandwidth slightly, and to count in the guard bands associated with telephone channels when speaking of the bandwidth. Modern telephone systems typically pass speech signals having a frequency range from about 300 to about 3400 Hertz. With the addition of guard bands, the total bandwidth of the channel is approximately 4000 Hertz, and this came to be called the "4 kHz channel." In the design of telephone systems, this has become the common denominator, and telephone systems are frequently defined in terms of so-and-so many channels capacity. If you were to come upon a 12 channel system, you would know that it had a bandwidth of 48 kHz (12 channels times 4 kHz per channel).

The concept of the 4 kHz channel is basic in multiplexing, and you will refer to it constantly. It is used to build larger clusters of channels, which will be introduced later in this chapter.

SPACE-DIVISION MULTIPLEXING

The simplest means of separating signals is to provide each with a separate path. This approach is known as space-division multiplexing, because you divided the available physical space into pieces, and apportioned a piece to each signal.

Suppose that you want to see two television shows, "Dawson's Creek" and "Nova." If they were to be sent into your living room using space division multiplexing, you would need two television sets, each connected to a separate input circuit. On one set, you could see "Dawson's Creek"; and on the other, "Nova." You could not see "Dawson's Creek" on the "Nova" set unless you physically switched the circuits connected to the inputs of the sets; the signal circuits between each program source and each set are completely separate. This is the principle of space-division multiplexing: a separate physical path for each signal.

Installation of space-division multiplexing is resource-intensive. A different physical circuit must be provided for every signal path. You are accustomed to this approach with many common systems. The multi-line telephone in an office has a separate pair of wires to connect each line on the phone to the telephone distributing frame in the office building. This is space-division multiplexing. Think of the electrical circuits in your home as signal paths. Each circuit has its own physical wire path from the fuse box to the outlet boxes. This is space division multiplexing. The voltmeter, oil pressure gauge, and fuel tank gauge in your auto are all connected to their dashboard instruments by separate circuits leading from the gauge to the sensor. Once again, space-division multiplexing.

Advantages and Disadvantages

Space-division multiplexing is simple and straightforward. It can be very economical over short physical distances because no special equipment is needed other than the wires carrying the circuits. Using separate circuits minimizes interference among signals being carried between the same two points. Circuits on parallel paths can have quite different electrical properties without influencing one another.

On the debit side of the ledger, space-division multiplexing can be very expensive, especially when a number of physical circuits must be provided over long parallel paths. Maintenance of space-division multiplex systems is often costly because of the number and length of circuits. Not insignificantly, space-division multiplexing is often wasteful of available bandwidth. A simple twisted pair cable can have a bandwidth of over 1 megahertz. When it is used for a single voice

circuit of only 4 kHz bandwidth, there is obviously a great deal of wasted capacity.

FREQUENCY-DIVISION MULTIPLEXING

Space-division multiplexing was used by the early telephone systems, and they rapidly ran out space for wires on poles. An Army officer, Major George O. Squier had an idea: "Why not separate these signals in frequency, rather than in space?" Since the wire transmission paths had a much wider bandwidth than the 4 kHz voice signals they were carrying, there was a great deal of expansion potential if signals could be "stacked up" in frequency and sent over the same pair of wires that were being used for just one conversation.

You are accustomed to frequency-division multiplexing in your everyday life. Suppose again that you want to watch either "Dawson's Creek" or "Nova", and that they are being televised at the same time. You merely select the proper channel on the TV set, and you can see the show you wish. Each channel is transmitted over a different band of frequencies, so both shows can be transmitted simultaneously without interfering with one another. You choose which one you want to watch by tuning to the proper frequency with the channel selector.

To understand frequency-division multiplexing (FDM), let's look at the diagram of a typical PDM telephone multiplexer, shown in Figure 18-1.

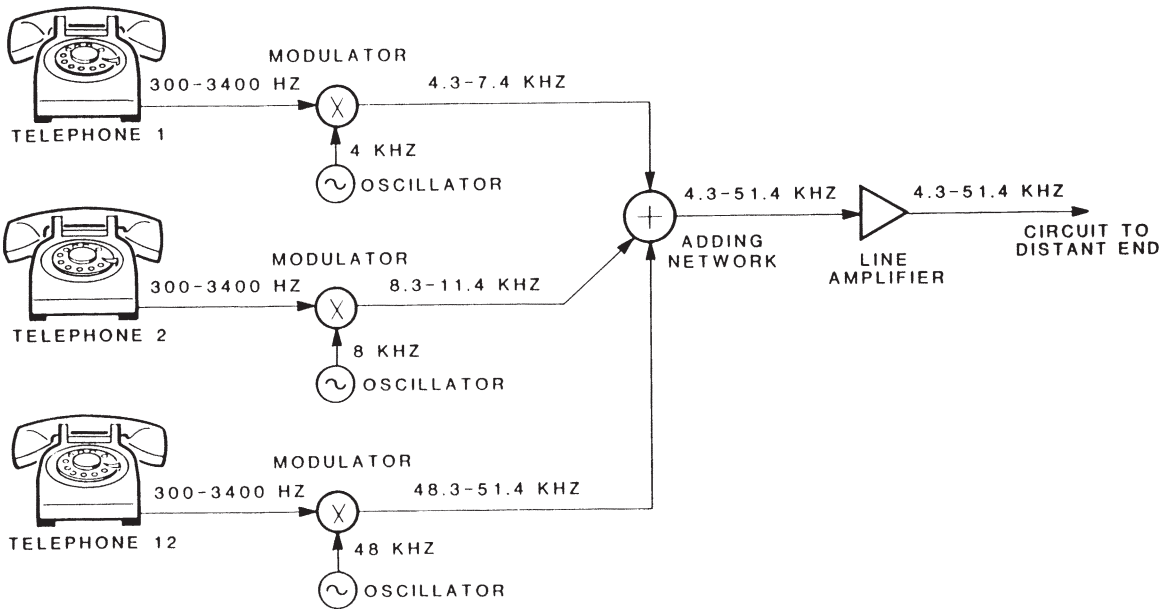


Figure 18-1

Block diagram of a frequency-division multiplexer.

At the inputs, you have twelve telephone channels, each carrying (or capable of carrying) an analog signal of 4 kHz bandwidth. We know that each input has a 4 kHz bandwidth because it is band-limited by filters before being applied to the multiplex channel input. Remember, the 4 kHz figure includes guard bands. The voice signal itself is usually limited to a band of about 300-3400 Hertz. Each of these signals modulates a carrier. The output of the modulation process is a series of signals having 4 kHz bandwidths, and spaced in the frequency spectrum at intervals of 4 kHz. Thus, all of these conversations can be transmitted simultaneously over the same pair of wires that previously could carry only one such conversation.

At the receiving end, a demodulator performs a mirror-image process to extract the original signals and return them to baseband where they can once more be heard by the human ear.

Figure 18-1 shows all of the inputs being multiplexed. In practice, an FDM system multiplexes all of the inputs to frequencies above baseband, and the baseband channel, that is "left over" as a result, is used as an order wire. An order wire is a channel for communications between the technicians that operate the multiplex equipment.

We said earlier that the 4 kHz channel was the building block for signal multiplexing. There are other, larger, building blocks as well. There is no reason to stop multiplexing at 12 channels. So long as the transmission medium has sufficient bandwidth, why not put together 24 channels or more?

No reason at all. However, the way things have developed, this is usually done in multiples of 12. This unit of 12 channels is referred to as a group. Each individual 4 kHz signal bandwidth is called a channel, and the equipment that accepts the 4 kHz baseband input is referred to as a channel pack. Figure 18-2 shows the concept of a high-capacity multiplexer. The outputs of five 12-channel groups are themselves multiplexed to form a supergroup of 60 channels. Supergroups are similarly multiplexed to form larger assemblies of channels in high capacity systems. A collection of 10 supergroups is commonly referred to as a mastergroup.

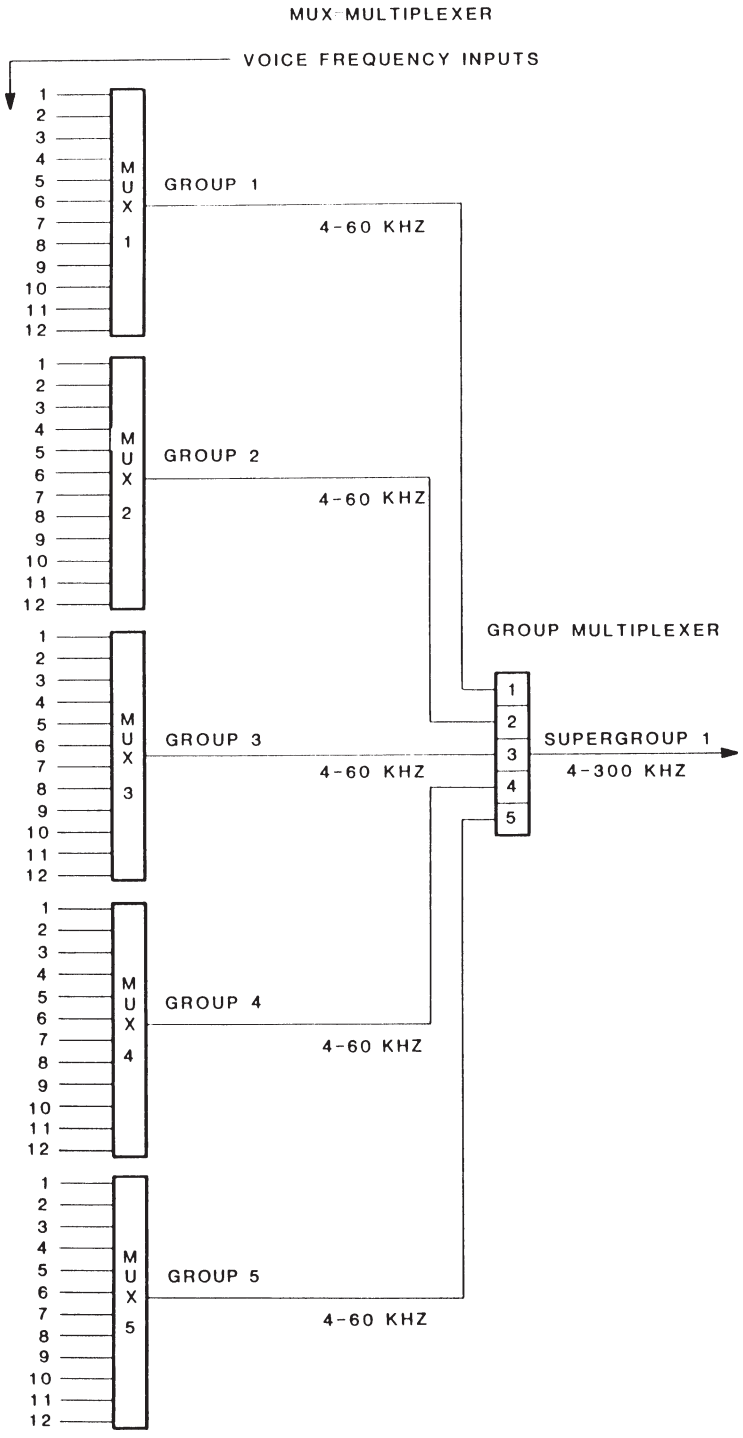


Figure 18-2

Forming an FDM supergroup.

The advantage of building FDM equipment in multiples of 12 and 5 is that it permits modularity. The frequencies used for multiplexing are not standard in the industry, but they are standard among the products of a given manufacturer. FDM systems are built by simply putting together the required number of channel packs, group banks, and supergroup banks. Using the concept of modularity, groups can even be dropped off and picked up along the way, as shown in Figure 18-3.

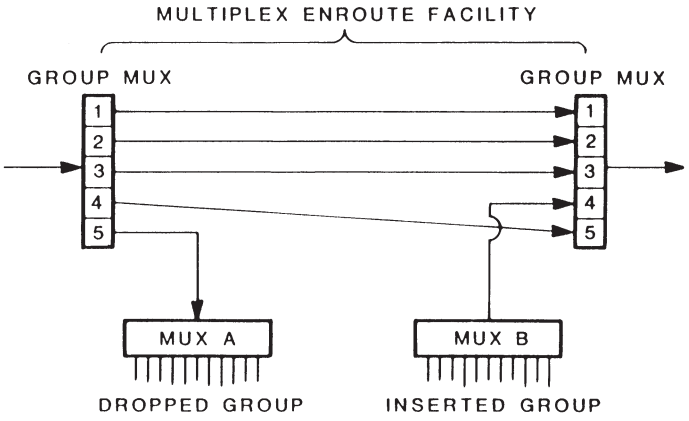


Figure 18-3

Dropping and adding PDM groups.

FDM is widely used in the telephone systems of the world. However, there is incompatibility between systems. As you have noticed, it is not necessarily possible to interchange the equipment of different manufacturers, even from the same country. The increasing volume of international communications is forcing these interface issues to be identified and dealt with. There is currently a wide variety of multiplex interface schemes operating at group levels and above. The ultimate interface is to demodulate everything to baseband and then remultiplex it. This is a brute-force approach, but it works. It is often used to interconnect incompatible equipment, rather than investing in sophisticated interface units.

Advantages and Disadvantages

The primary advantage of FDM is that it extends the capacity of the communications plant. The same wires, or cables, that before, could carry only a single signal, can carry hundreds of signals using FDM technology. It is much less expensive to install FDM on a long route, than it is to construct a separate physical circuit that parallels one already in place.

FDM also allows all signals to be adjusted to, more or less, standard levels. Weak signals can be amplified at the channel level, and all signals can be amplified, periodically, at the group or supergroup level, to permit transmission of signals over long distances. Amplifiers can be used with space-division multiplex, too, but the costs of separate amplifiers for each of the circuits add up quickly.

There are also disadvantages to FDM. FDM demands stable multiplex carrier frequencies, and requires the frequencies at both ends to be synchronized with one another. If they are not quite synchronized, there will be frequency and phase distortion.

Using FDM means that digital signals have to be reduced to analog form for transmission. Thus, we lose the advantages of pulse regeneration in digital transmission, and have to deal with noise encroaching on our signal, just like any other analog signal.

FDM systems are primarily designed to carry human conversations, and have been carefully designed around the statistics of the human voice. One of those characteristics is a low average power level. Digital signals have a high average power level, and this tends to overload the FDM channels. In fact, several digital signals on an FDM group, or supergroup, can overload the entire group. The effect of overloading is distortion, which can cause the contents of one channel to "splatter" into adjacent channels. Trying to sort out the signal at the receiver can be a very unrewarding experience.

The amplifiers in FDM systems do not generally have flat responses over their passbands. This means that some frequencies are amplified more than others, and this amplitude distortion can make the detection of digital signals difficult.

Any filter introduces non-linear phase variations over its passband, especially near the cutoff frequencies. The filters used to band-limit the input signals to the channel packs are no exception, and the phase distortion that they introduce on the transmission path complicates the problems of detecting digital signals without error.

TIME-DIVISION MULTIPLEXING

So far, you have separated signals in space and frequency to get them from one place to another. We can also separate them in time.

Once again, consider our two TV shows: "Dawson's Creek" and "Nova." If they are both to be shown on the same channel, then the only choice is to show them at separate times, otherwise they will interfere with one another, and you will not be able to sort out the result. This is time division multiplexing. The signals are being sent over the same transmission path, at the same frequency, but at different times. By knowing the time at which each signal appears, you can view the received program and it will be understandable.

Figure 18-4 shows the concept of a time-division multiplexer. For the moment, let's assume the input signals are pulse trains with known pulse rates.

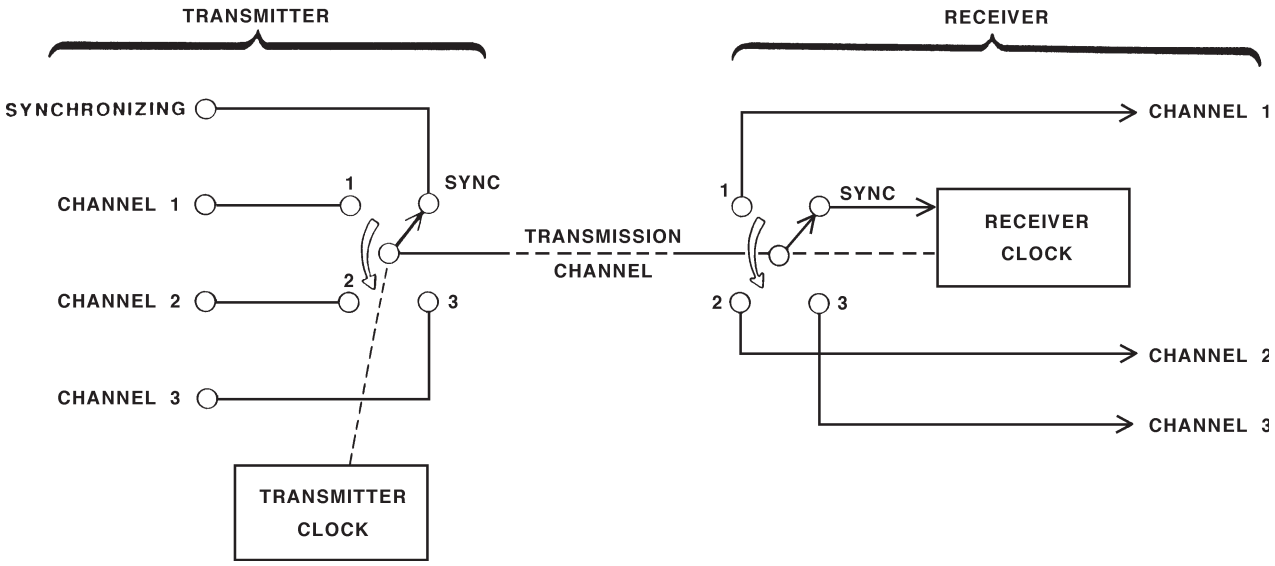


Figure 18-4

A very simplified diagram of a TDM transmitter and receiver.

Looking at Figure 18-4, you see that the rotating switch acts just like the distributor on an automobile engine. It connects each input to the transmission line for a specified period of time. With the correct synchronization between the rotating switch and the input pulses, each input is connected to the transmission line, in turn. The result is an output pulse stream that has a pulse rate equal to the number of channels times the channel data rate. Therefore, the output pulse rate is 12 times greater than each of the input channels for a twelve channel system.

At the receiving terminal, the process is reversed. The rotating switch is connected, in turn, to each of the output lines. This routes the pulses from their positions in the consolidated pulse stream, into individual circuits. There, each pulse stream looks like a replica of the input to that circuit at the transmitter.

You can easily see that the key to success of this scheme is synchronization of the rotating switches. If either one gets out of step, pulses will be misrouted into the wrong pulse stream, and the outputs quickly become gibberish.

In actual practice, rotating switches are seldom used (although they can be found in certain radar and industrial applications). Instead, digital circuitry is used to perform the function of the rotating switch in the diagram.

Since it is a purely digital signal, PCM is ideal for the application of TDM techniques. Figure 18-5 shows the basis for constructing a TDM signal using PCM applied to speech. Each channel must be sampled 8000 times per second, but each does not have to be sampled at precisely the same moment. By taking the samples at successively later times for each channel, you can build a signal segment that contains a sample from each of the channels. When you once again return to sample Channel 1, a new segment of the multiplexed signal begins. This signal segment is called a frame.

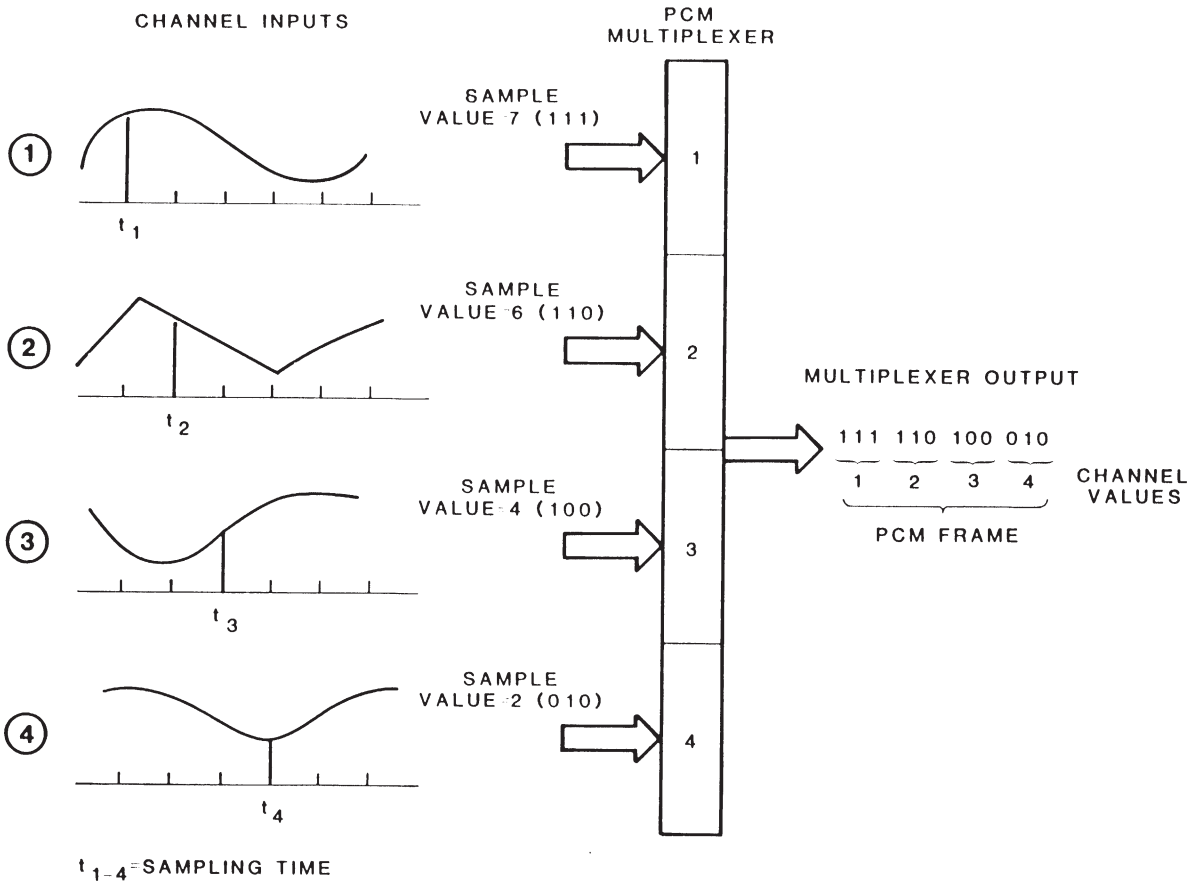


Figure 18-5

Simplified building of a PCM frame.

A TDM-PCM bit stream consists of many frames, one after the other. In order to permit the receiver to determine where it is in the sequence of the signal, a distinctive frame pulse or sequence of pulses is generally transmitted to mark the beginning of each frame. Even if the frame pulse is distorted or not received, the result is only the loss of a single frame, which is not a very large part of any one of the channel samples. The overall effect on fidelity of losing a single frame pulse is

not severe in voice transmission. By synchronizing the PCM system with the frame pulses, the system re-synchronizes every frame. If a frame pulse is lost, re-synchronization occurs with the next frame pulse received, and the loss of a single pulse does not cause irretrievable errors by destroying system synchronization.

PCM is most commonly carried on time division multiplexed systems, and when one speaks of PCM, TDM is inferred. However, TDM does not require PCM inputs. Provided Nyquist's criteria for sampling rate is observed, you can sample analog inputs directly at the modulator input, and produce PAM or PPM pulses to be sent down the transmission path. There are some applications that do this.

Similarly, PCM does not require TDM. It might be necessary to transmit PCM encoded voice on a path over which FDM is operating. This can be done in the same manner that any digital signal is sent over an analog network; translate it into an analog signal that the network can handle. Much of the benefit of PCM is lost when this is done, but the point is that PCM signals can be sent by any means necessary.

Practical TDM-PCM Systems

The first widely-used commercial TDM-PCM transmission system was the Bell System T1 carrier system. The T1 system is a PCM system with an output bit rate of 1.544 million bits-per-second. T1 carrier equipment is interconnected by ordinary wire pairs with regenerative repeaters spaced just over a mile apart along the wires. This system carries 24 voice channels. Each channel is sampled at an 8 kHz rate, and seven bits are used to encode the signal level. An eighth bit is used to carry supervisory signalling data for the channel. Each frame contains a framing bit, and a special sequence of bits is used in conjunction with the framing bit to signify the beginning of each frame. With 24 channels, you can see that each frame contains 193 bits:

$$24 \text{ channels} \times 8 \text{ bits/channel} = 192 \text{ bits} + 1 \text{ frame bit} = 193.$$

The T1 carrier system is only the first in a hierarchy of PCM multiplexers which are modular, like the building blocks of FDM systems. The next larger T2 carrier operates at a rate of 6.3 million bits-per second. You can be think of a T2 carrier as a group modulator accepting the inputs from four T1 carriers. This form of multiplexing has become so economical that in many localities, it is cheaper to install a T1 or T2 carrier to multiplex many individual telephone lines between the telephone instrument and the central office than it is to install separate wire pairs for each phone!

Advantages and Disadvantages

TDM offers the cost effectiveness of LSI digital circuitry, and also the ability to minimize circuit noise through signal regeneration. When TDM is combined with PCM, it is possible to achieve very quiet, efficient circuits over long distances.

The basic disadvantage of TDM is its complexity. While the circuitry is largely integrated, and entire modules can be easily replaced when they fail, it is first necessary to find what failed. This requires skilled technicians and sophisticated test equipment.

SUMMARY

In this chapter you applied what you learned to multiplexing. You learned that many signals can be sent over the same route without interfering with one another if they are separated in space, frequency, or time. You also found that TDM, coupled with PCM, provides an excellent, low-noise, and relatively inexpensive long distance communications medium.

Multiplexing has become the rule rather than the exception in communications systems. There is a wide variety of multiplexers in use in the communications industry. Given the growth of data communications, the investment in installed equipment, and the wide-ranging nature of new communications applications, it is likely that you will encounter most of the multiplex types you have studied at some point in your communications work.

SELF-TEST REVIEW

1. What is multiplexing? _____

   _____

   _____

2. What are the two basic types of multiplexing? _____

   _____

   _____

3. What is TDM? _____

_____

_____

4. A complete set of TDM pulses from one sync pulse to the next is called a _____ .

5. Switching between channels in a TDM transmitter and receiver is controlled by the _____ and the _____, respectively.

6. What is FDM? _____

_____

_____

7. What is the primary disadvantage of FDM? _____

_____

_____

Section 6

ANTENNAS

CHAPTER 19 — Transmission Lines

INTRODUCTION

Antennas are what make radio communications "wireless." An antenna converts radio frequency current into electromagnetic waves for transmission, or converts electromagnetic waves into radio frequency current for reception.

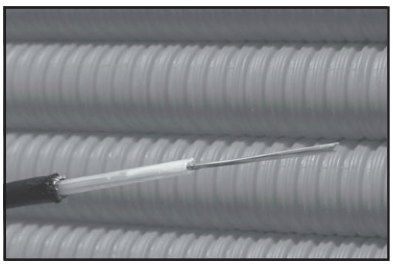
The three main elements of an antenna system are:

1. The transmission line that conveys energy to and from the antenna.

2. The antenna itself.

3. The path the radio wave takes between the transmitter and receiver antennas.

This chapter describes transmission lines. The primary purpose of a transmission line is to transfer energy from a source to the load. An example is the transmission line used to transfer RF energy from a transmitter to the antenna. Lines are also used to interconnect equipment such as transmitter to power amplifier, receiver to speaker, etc. When the energy being transferred is DC or audio, the lines present no special problems. At these frequencies, transmission line characteristics can be ignored. However, at higher frequencies these line characteristics cannot be ignored. These characteristics are related to the signal's wavelength, which is the distance the wave travels in one cycle. This section describes these characteristics in detail.



OBJECTIVES

When you complete this chapter, you will be able to:

- State the primary purpose of a transmission line.
- State the difference between balanced and unbalanced transmission lines.
- State what determines the characteristic impedance of a transmission line.
- Write the equations for standing-wave ratio.
- Define reflectometer.
- State the characteristics of resonant transmission lines.
- Describe the comparative advantages of coaxial cable and waveguide transmission.

THE BALANCED TRANSMISSION LINE

Most of us are familiar with the "TV twin lead" type of transmission line shown in Figure 19-1A. This is the type of lead-in wire used to connect an antenna to a television receiver. This type of transmission line is known as a **balanced** line. Both wires of the line carry RF current and the current in each wire is 180° out of phase with the other wire. (See Figure 19-1B.) Therefore, the current in this type of transmission line is balanced with respect to electrical ground. On the other hand, with an unbalanced transmission line, one wire is at ground potential, while the other wire carries the RF current. An example of an unbalanced line is the coaxial cable used in cable television hookups. This is shown in Figure 19-1C.

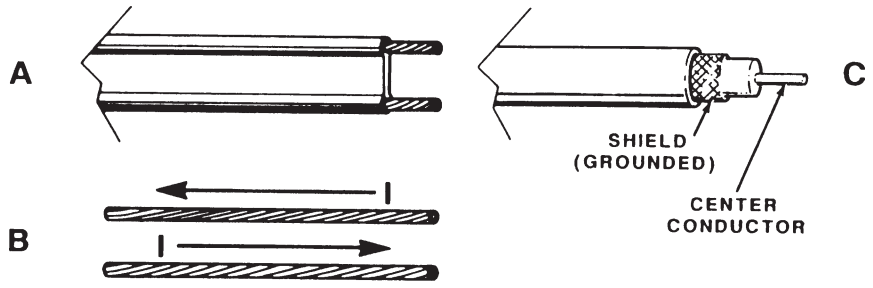


Figure 19-1
Transmission lines.

The equivalent circuit of a balanced transmission line is shown in Figure 19-2. Inductance is present because whenever current flows through a conductor, a magnetic field is produced around it. Since a potential difference exists between the two conductors of the transmission line, due to the 180° phase difference, an electrostatic field exists between the conductors. Therefore, some capacitance also exists. The equivalent circuit shown in Figure 19-2 represents the ideal transmission line. In a practical line, a series resistive element is present due to the resistance of the conductors. Also, a parallel resistive element is present due to leakage in the dielectric material that separates the two conductors. The inductive, capacitive, and resistive properties of the transmission line present a specific impedance to any AC waveform.

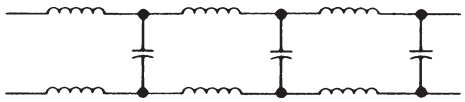


Figure 19-2
Balanced line equivalent circuit.

If an RF generator is connected to an infinitely long transmission line, a forward traveling wave of RF energy will start at the generator and travel down the line. This is shown in Figure 19-3A. This applied RF energy causes a current to flow in the transmission line and a resulting potential difference, or voltage, to appear across the two conductors. If you now take current and voltage measurements, with an AC meter, at several points on this line, you will find that the amplitude of voltage and current is constant at all points. That is, the applied RF energy causes a specific current and voltage to appear at all points on the transmission line. A graph of voltage and current on this transmission line is shown in Figure 19-3B.

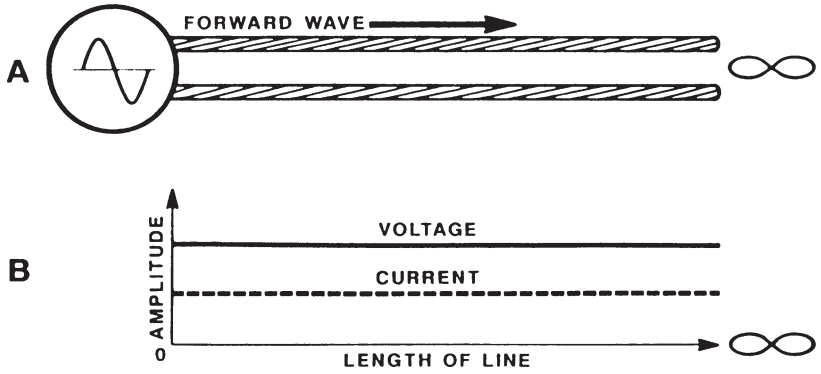


Figure 19-3
An infinite transmission line.

Since voltage (E) and current (I) are constant at all points, impedance (Z) is also constant, proven by the Ohm's Law equation. Therefore, the infinite length of transmission line has a constant impedance. This is known as its *characteristic impedance*. Every transmission line has a specific characteristic impedance which is due to its physical construction. The equation for the **characteristic impedance** of a balanced transmission line where the conductors are insulated by air is:

$$Z_o = 276 \log \frac{d}{r}$$

where Z_o = characteristic impedance in ohms

d = characteristic impedance in ohms
separation of conductors

r = radius of conductors (using the same unit of measurement as with d)

Figure 19-4 shows a balanced transmission line with measurements d and r indicated.

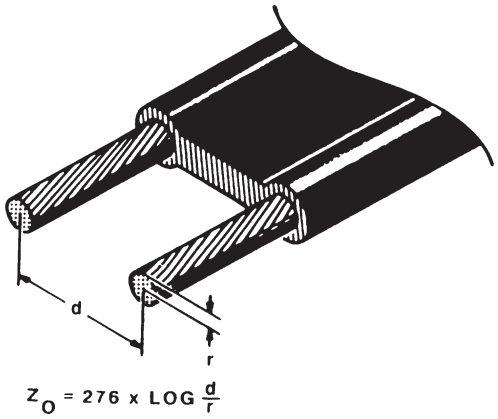


Figure 19-4
Balanced transmission line dimensions.

The following example shows how to find the characteristic impedance of a typical balanced transmission line. The distance between conductors is 1.22 cm and the conductor radius is 0.1 cm. Using the equation for characteristic impedance:

$$Z_o = 276 \log \frac{d}{r}$$

$$Z_o = 276 \log \frac{1.22}{0.1}$$

$$Z_o = 276 \times \log 12.2$$

$$Z_o = 300$$

Thus, the characteristic impedance of this particular line is 300 ohms. If, as with common TV twin-lead, the insulator is not air, the characteristic impedance is reduced. Also, the loss attributable to the line increases.

If you connect an infinitely long length of this 300 ohm line to an RF generator, the input impedance that the generator sees is 300 ohms. In fact, the impedance at all points on the line is 300 ohms. Therefore, if you cut the line to some practical length and connect a 300 ohm resistor across the far end of the line, the generator still sees a 300 ohm input impedance. You are, in essence, replacing the lost portion of the

line with a 300 ohm resistor. Since the resistor has the same impedance as the lost portion of the line, the previous conditions are not upset; the transmission line still behaves as though it were infinitely long. The forward traveling wave of RF energy travels the length of the line and is completely dissipated in the 300 ohm resistor. This condition exists whenever a transmission line is correctly terminated; that is, whenever the load impedance matches the characteristic impedance of the transmission line. In the example given, the 300 ohm load resistor matched the 300 ohm transmission line.

STANDING WAVES

If a transmission line has infinite length or is terminated in its characteristic impedance, all of the power applied to the line at one end is absorbed or dissipated by the load at the other end. However, if a transmission line is terminated in an impedance other than its characteristic impedance, not all of the applied power is absorbed by the load. The remaining power is **reflected**.

When a transmission line is incorrectly terminated, the power that is not absorbed by the load is sent back toward the generator. The amount of this reflected power is directly proportional to the amount of impedance mismatch at the load. For example, if the line is terminated by a short circuit or an open circuit, none of the power is dissipated in the termination and all of it is reflected back to the generator. As the termination impedance approaches the line's characteristic impedance, less and less power is reflected, until finally, when the impedances match, no power is reflected.

When power is applied to a transmission line, a voltage and current appear whose values depend upon the characteristic impedance (Z_o) and the applied power. The voltage and current waves travel to the load (Z_L) and, if $Z_L = Z_o$, the load absorbs all of the power and none is reflected. The only waves present at this time are the forward traveling waves of current and voltage from the generator to the load.

However, if Z_L is not equal to Z_o, some power is absorbed and the rest is reflected. Therefore, one set of current and voltage waves is traveling toward the load and another set is traveling back to the generator.

This is shown in Figure 19-5. These two sets of traveling waves, going in opposite directions, alternately add to and subtract from one an other due to their relative phase differences. This, in turn, sets up a pattern known as **standing waves** along the line. The standing wave pattern for an open-circuited line is shown in Figure 19-6. You can see that **stationary** voltage and current minimums and maximums have appeared. Compare this voltage and current pattern to that of the infinite length transmission line shown in Figure 19-3. In the infinite or correctly terminated line, there are no reflected waves to interfere with the forward waves. How do these standing waves develop?

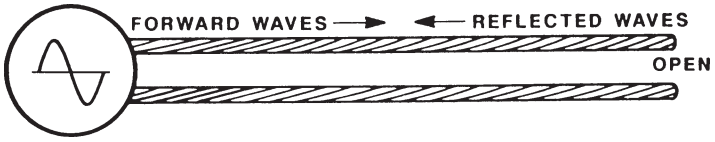


Figure 19-5
Forward and reflected waves.

For the moment, consider only the forward traveling voltage and current waves. At the load, the current is zero and the voltage is maximum because the load is an open circuit. This is shown in Figure 19-6. The reflection that occurs at the open circuit affects both voltage and current. The current now starts traveling back to the generator, with a 180° phase inversion, but the voltage is unchanged in phase. Therefore, at a point exactly a quarter-wavelength (1/4λ) from the load, the voltage is permanently zero, as shown in Figure 19-6, because the forward and reflected voltage waves are exactly 180° out of phase. This occurs because the reflected wave has had to travel a distance of 1/4λ ÷ 1/4λ or 1/2λ farther than the forward wave. Since 1/2λ is equal to a phase change of 180°, the two waves cancel and voltage is zero at this point. The current wave has also traveled an extra 1/2λ; however, since it underwent a 180° phase inversion on reflection, its total phase change is 360°. Therefore, because the forward and reflected current waves are in phase, reinforcement takes place, and a current maximum occurs at exactly the same point as the voltage minimum.

Figure 19-6 shows that at 1/2λ from the load, there is a current minimum (zero in this case) and a voltage maximum. This occurs because the forward and reverse voltage waves are now in phase, voltage has traveled a distance of one wavelength (360°) to return to this point. At this same place, the current waves cancel, due to the additional 180° phase inversion that occurred on reflection. Figure 19-6 shows that these maximum and minimum points repeat every half-wavelength and reverse every quarter-wavelength.

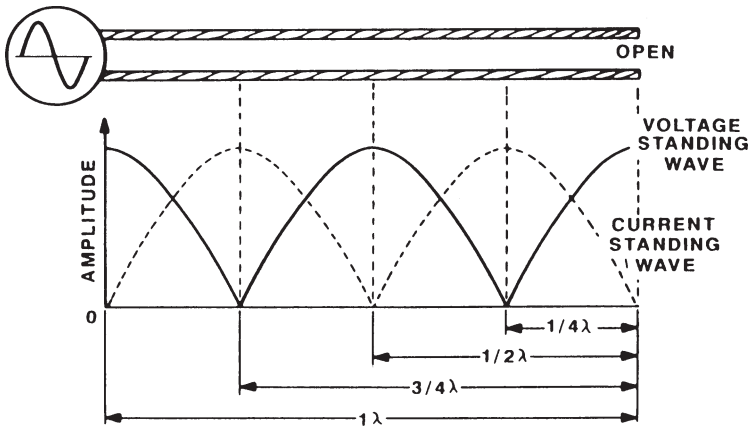


Figure 19-6
Standing wave pattern for an open circuit line.

Note that this condition is permanent and is determined by the load. Figure 19-7 shows the standing-wave pattern for a short-circuited transmission line. Here, the current at the load is maximum and the voltage is zero because the load is a short circuit. At a quarter-wavelength from the load, current is zero and voltage is maximum. At a half wavelength from the load, current is maximum and voltage is zero. Thus, the standing-wave pattern for a short-circuited line is just the opposite of that for an open-circuited line. This shows that the load does indeed determine the standing-wave pattern.

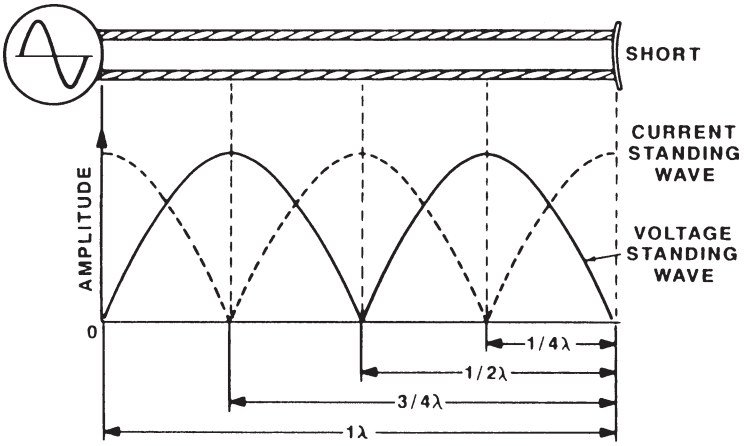


Figure 19-7
Standing wave pattern for a short-circuited line.

STANDING WAVE RATIO

The ratio of maximum current to minimum current or maximum voltage to minimum voltage along a transmission line is called the standing-wave ratio (SWR). This ratio is a measure of the mismatch between the load and the transmission line. For example, the SWR is equal to 1:1 when the load is perfectly matched, and equal to when the load is a short or open circuit. If the transmission line is terminated in a purely resistive load, you can use the following equation to find the standing-wave ratio:

$$SWR = \frac{Z_o}{R_L} \text{ or } SWR = \frac{R_L}{Z_o}$$

(Whichever is greater than 1)

Where R_L = Load resistance

The standing-wave ratio is always equal to or greater than one. Regardless of whether R_L is half as large or twice as large as Z_o the ratio of a voltage maximum to a voltage minimum is 2:1. Therefore, the degree of mismatch is the same in both instances.

The higher the SWR, the greater is the mismatch between line and load. Also, in practical transmission lines, power loss increases with SWR. Therefore, a low value of SWR is always sought, except when the line is used as a tuned or resonant circuit.

TRANSMISSION LINE RESONANT CIRCUITS

A transmission line that is terminated in its characteristic impedance has an SWR of 1:1 and is called a **non-resonant** or flat transmission line. It is non-resonant because it presents a constant load impedance to the generator regardless of the operating frequency. On the other hand, a transmission line with an SWR greater than 1:1 presents a variable load impedance that is dependent upon the operating frequency. Therefore, a line with an SWR greater than 1:1 is called a **resonant** transmission line.

Figure 19-8 shows a 25 cm length of transmission line that is terminated with a short circuit. Therefore, the SWR is infinite. Figure 19-8A shows the standing-wave pattern that develops on the line when a 300 MHz signal is applied. The 25 cm length of line is 1/4λ at 300 MHz because:

$$\lambda(\text{meters}) = \frac{3 \times 10^8 \text{meters/second (velocity of waves)}}{\text{frequency } (H_Z)}$$

$$= \frac{3 \times 10^8}{3 \times 10^8} = 1 \text{ meter}$$

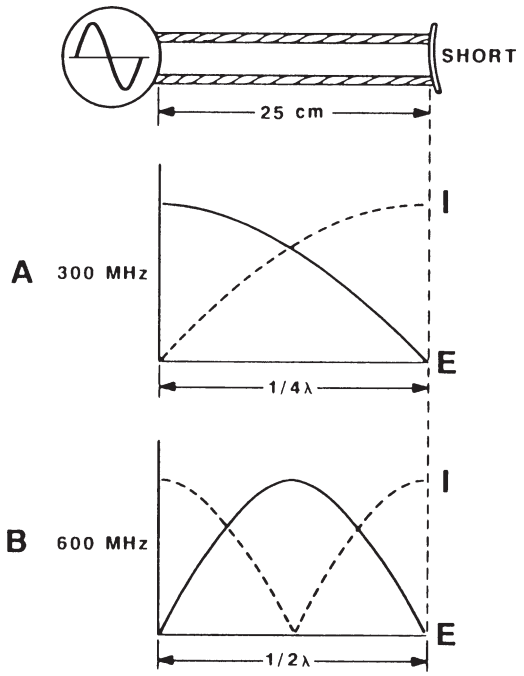


Figure 19-8
A resonant transmission line.

Therefore, 1/4λ is 0.25 m, or 25 cm at 300 MHz. Notice that the input of the line in Figure 19-8A has maximum voltage and minimum, or zero, current. This corresponds to an open circuit and it is in fact the input impedance of the 1/4λ shorted transmission line. The quarter wave line actually inverted the impedance of the load. If the load had been all open circuit, the input impedance would have been a short circuit. Since these conditions are only true when the line is exactly 1/4λ long, if the input frequency changes, so does the input impedance. Because the transmission line is frequency dependent, it is said to be resonant. In fact, the 1/4λ shorted line acts exactly like a parallel resonant circuit. That is, at its resonant frequency or the frequency at which it is exactly 1/4λ long, its input impedance is infinite, or open. At all other frequencies, its input impedance becomes either capacitive or inductive depending upon whether it is above or below its resonant frequency.

Transmission lines are used as both parallel and series resonant circuits in VHF, UHF, and microwave equipment They not only offer an extremely high Q; but, are often the only practical method of obtaining a resonant circuit at these frequencies. This is because, as frequency increases, the inductors and capacitors required become smaller and smaller. When this happens, circuit Q drops and bandwidth increases. The only way to improve this is to use transmission line resonant circuits.

To be sure you understand this important concept, consider Figure 19-8B. Here is the same 25 cm length of transmission line we described earlier. However, since the input frequency has been doubled to 600 MHz, the line appears as a half-wavelength. This accounts for the different standing-wave pattern. Note that the input now has maximum current and minimum, or zero, voltage. Therefore, the input is a short circuit for a 1/2λ shorted transmission line. Since this occurs only at the frequency at which the line is exactly 1/2λ, it acts as a resonant circuit. In fact, the 1/2λ shorted line acts like a series resonant circuit. Notice that the 1/2λ line duplicates the terminating impedance at its input. Therefore, the input of a 1/2λ open transmission line will also be an open.

To summarize, a 1/4λ line inverts its terminating impedance. Therefore, the input of a shorted 1/4λ line is an open, and the input of an open 1/4λ line is a short. Thus, a shorted 1/4λ can be used as a parallel resonant circuit, and an open 1/4λ line can be used as a series resonant circuit. Conversely, a 1/2λ line duplicates its terminating impedance. Therefore, the input of a shorted 1/2λ line is a short and, the input of an open 1/2λ line is an open. Thus, a shorted 1/2λ line can be used as a series resonant circuit, and an open 1/2λ line can be used as a parallel resonant circuit.

A correctly terminated transmission line has a constant input impedance regardless of frequency or line length. For example, if a 300 ohm transmission line is terminated by a 300 ohm resistor, its input impedance will be 300 ohms no matter what the frequency or the length of the line. It is for this reason that a matched condition, indicated by an SWR of 1, is always sought when installing a transmission line. The only time a high SWR is desired is when the line is being used as a resonant circuit as shown previously.

COAXIAL TRANSMISSION LINES

Figure 19-9A shows a coaxial type of transmission line, which is one of the most commonly used types of lines. It is an unbalanced line, since it consists of a shield that is at ground potential, and a center conductor. The center conductor "carries" the RF energy and the shield prevents this energy from being radiated into space. Thus, losses due to radiation are kept to a minimum. With a balanced line, radiation is minimal due to the balanced currents. The radiated fields created by each line are 180° out of phase and, therefore, cancel.

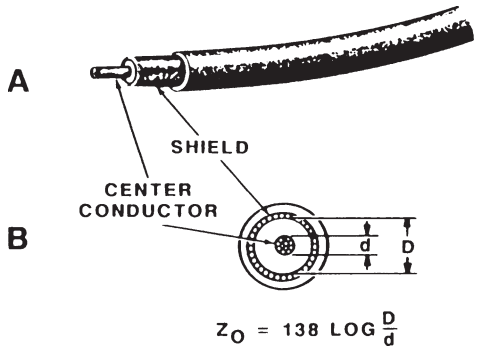


Figure 19-9
Coaxial transmission line.

A coaxial line has a specific characteristic impedance just as a balanced line does. The equation for the characteristic impedance of a coaxial line with air insulating the shield from the center conductor is:

$$Z_o = 138 \times \log \frac{D}{d}$$

Where Z_o = characteristic impedance in ohms.

D = inside diameter of shield.
d = outside diameter of the center conductor.

These dimensions are shown in Figure 19-9B.

As with balanced lines, using an insulation other than air reduces the characteristic impedance and increases line loss.

COMPARISON OF COAXIAL AND BALANCED LINES

When transmission line characteristics were described earlier, the lines were assumed to be perfect. However, practical transmission lines exhibit power losses. These losses are due to the resistance of the conductors, leakage in the dielectric material that separates the conductors, and in some cases, radiation. Figure 19-10 shows the normal attenuation of three typical transmission lines versus the operating frequency. RG-58/U cable is a small diameter 52 ohm coaxial cable, RG-8/U cable is a larger diameter 52 ohm coaxial cable, and 214-022 cable is a 300 ohm balanced transmission line. You can see that the balanced transmission line offers the lowest attenuation; however, this advantage is offset to some degree by its installation requirements.

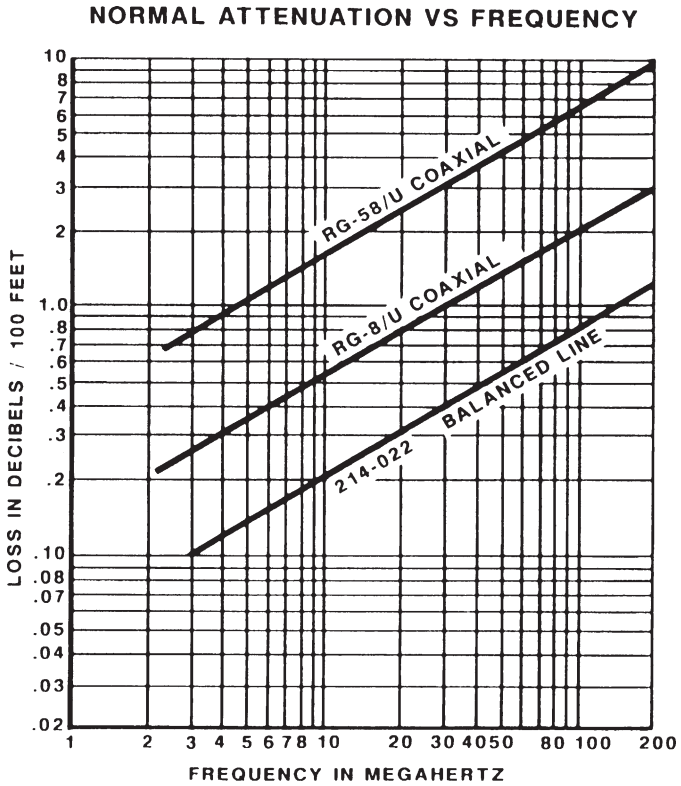


Figure 19-10
Attenuation in transmission lines.

When you install a balanced transmission line, you must prevent it from collecting moisture. Any moisture that accumulates on the line will change the dielectric between the two lines, which changes its characteristic impedance. The result is reflections and additional line losses. You must also keep the line away from other conductors such as downspouts, rain gutters, and metal towers. The line must be spaced at least two to three times the conductor separation from any other conductors. For example, if the distance between the two wires of the transmission line is 1/2λ, the line must be spaced 1λ to 1-1/2λ away from other conductors. This requires standoff insulators.

Coaxial line, on the other hand, has an outside shield that is at ground potential, so you can mount it anywhere. You can tape it directly to any metal object without adverse effects. Although moisture can affect coaxial cable if it enters the cable, it has little effect when proper care is taken to prevent moisture buildup on the inside of the cable. Thus, while balanced line offers the lowest attenuation, coaxial cable is widely used because it's easy to install.

In addition, the balance of a balanced line is easily disturbed by such practical items as mounting brackets, moisture, etc. These imbalances cause increased radiation from the line, and also make the line susceptible to spurious signals being induced on it. The shield of a coaxial cable significantly reduces radiation and spurious signals, another reason why coaxial lines are widely used.

SPECIALIZED COAXIAL CABLES

In an effort to reduce signal losses at high frequencies where long cable runs are required, a special type of coaxial cable called "hard line" or "Heliax®" is used. This type of coaxial cable is available with a foam dielectric or an air dielectric and requires special connectors.

Foam dielectric hard line (see Figure 19-11) is available in six standard sizes from 1/4" to 1-5/8" . The amount of signal loss exhibited by these cables is dependent upon the frequency, length, and cable size. At 450 MHz, for example, 100' feet of 1/4" foam dielectric hard line has a loss of about 4 dB, while 100' of 1-5/8" foam dielectric hard line has a loss of about 0.5 dB. As frequency increases, so does the loss. Frequency and cable size determine how much power the cable can handle.

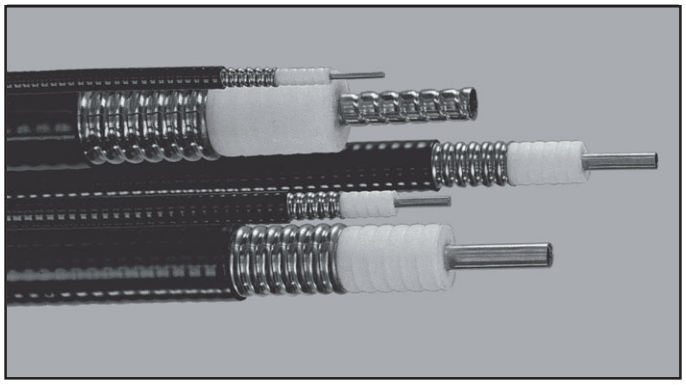


Figure 19-11
Foam dielectric hard line.
Photo courtesy Andrew Corp.

Air dielectric hardline (see Figure 19-12) is available in eight standard sizes from 1/2″ to 5″. This type of cable uses a thin helical-wound insulator to keep the center conductor a constant distance from the outer conductor. At 450 MHz, 100′ feet of 1/2″ air dielectric hardline has a loss of less than 2 dB, while 100′ of 5″ air dielectric hardline has a loss of about 0.2 dB. Since this type of cable does not have a solid dielectric, it is possible for condensation to build up inside the cable. To prevent this from happening, the cable is pressurized with a drying agent such as nitrogen.

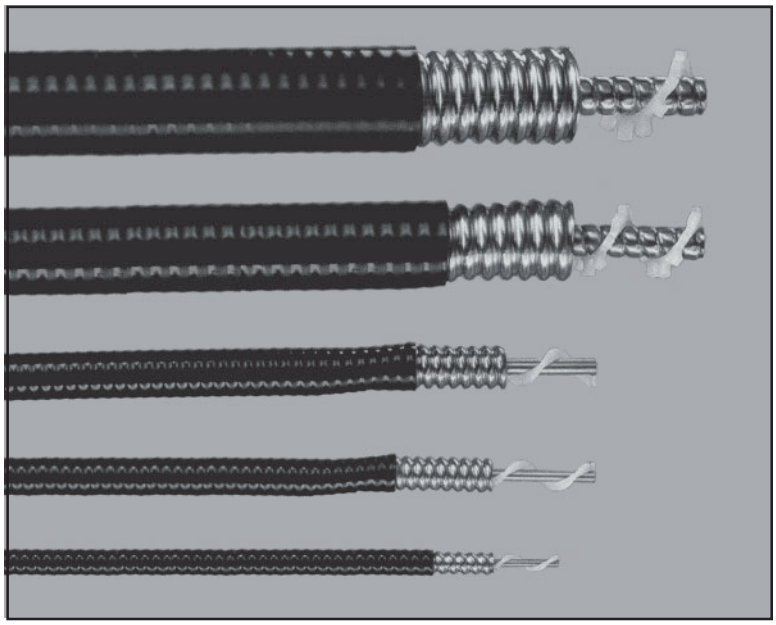


Figure 19-12
Air dielectric hard line.
Photo courtesy Andrew Corp.

WAVEGUIDES

At very high frequencies, the losses of practical transmission lines become significant. In an attempt to minimize losses and also provide practical, low-radiation transmission line, waveguides were developed.

A waveguide, shown in Figure 19-13, is merely a hollow metal conductor, usually rectangular or round, which transfers electrical energy from one place to another. Without going into the mathematical details, let' s say that the name of the device is descriptive. A waveguide guides electromagnetic waves from place to place, much as a coaxial cable does. It is merely another form of transmission line. It has very low loss and wide bandwidth, and so, it is suitable for carrying UHF or microwave signals.

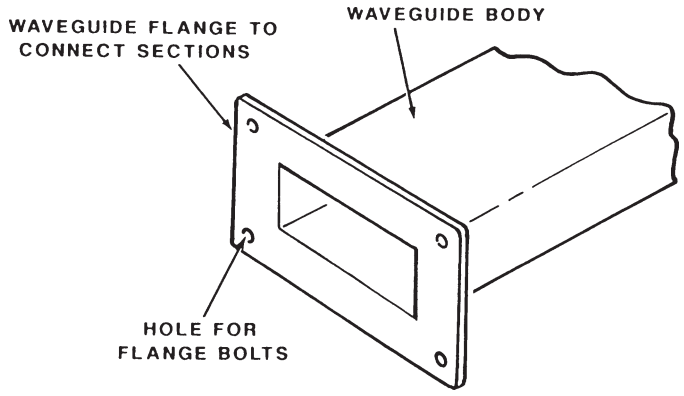


Figure 19-13
A rectangular waveguide.

The horizontal dimension of a rectangular waveguide generally defines the lowest frequency for which the waveguide is usable. That dimension is one-half wavelength at the lowest useful frequency, or cutoff frequency. So, if we were to build a waveguide to operate at 300 MHz, which has a wavelength of 1 meter, the waveguide would need to be half a meter wide. That is too large for most practical applications. However, at 3 GHz, the wavelength of the signal is only one tenth of a meter, and the waveguide at this frequency needs to be only one-half that in width, or one-twentieth of a meter (5 cm). That is only about two inches, and is a very practical size.

Because of its low loss, a waveguide is widely used as a transmission line at microwave frequencies. The natural shielding provided by the metal enclosure offers excellent isolation of the signal in the waveguide from signals outside the waveguide.

SWR MEASUREMENT

Standing-wave ratio is measured by a reflectometer, also called an SWR meter. A typical SWR meter is shown in Figure 19-14. This unit measures the ratio between forward and reflected power and displays it on a calibrated meter.

Figure 19-14
Reflectometer or SWR meter.

Figure 19-15 shows that the SWR meter is connected in series in the transmission line, between the transmitter and the load or antenna. SWR is measured by first tuning the transmitter for maximum output. The SWR sensitivity control is then set for full-scale meter deflection. The meter is actually measuring forward power wave voltage, when the SWR sensitivity control is set for full scale. Then, the meter is switched to measure reflected power. The amount of reflected power is indicated by the meter; however, the meter is actually calibrated to display the standing-wave ratio directly. This is a quick and easy method of checking SWR with a relatively low-cost instrument.

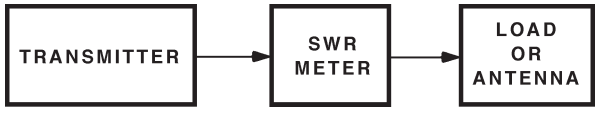


Figure 19-15
An SWR meter is connected in series in the line between transmitter and load.

A meter that measures the actual amount of reflected and forward power is shown in Figure 19-16. It is a directional wattmeter and, with the flick of a switch, you can measure forward and reflected power. A perfectly matched condition, SWR = 1:1, is indicated by zero reflected power. For other values of reflected power, the standing wave ratio is determined from the chart shown in Figure 19-15. You can use this chart to find the SWR when forward and reflected power are known. For example, if the forward power is 5 watts and reflected power is 0.2 watts, find these lines on the chart; the point where they intersect indicates the standing-wave ratio. In this case, they intersect on the diagonal line that represents an SWR of 1.5.

Figure 19-16
A directional wattmeter.

The chart shown in Figure 19-17 reminds you that SWR is determined from the ratio of voltages or currents in the transmission line, not the ratio of forward and reflected power.

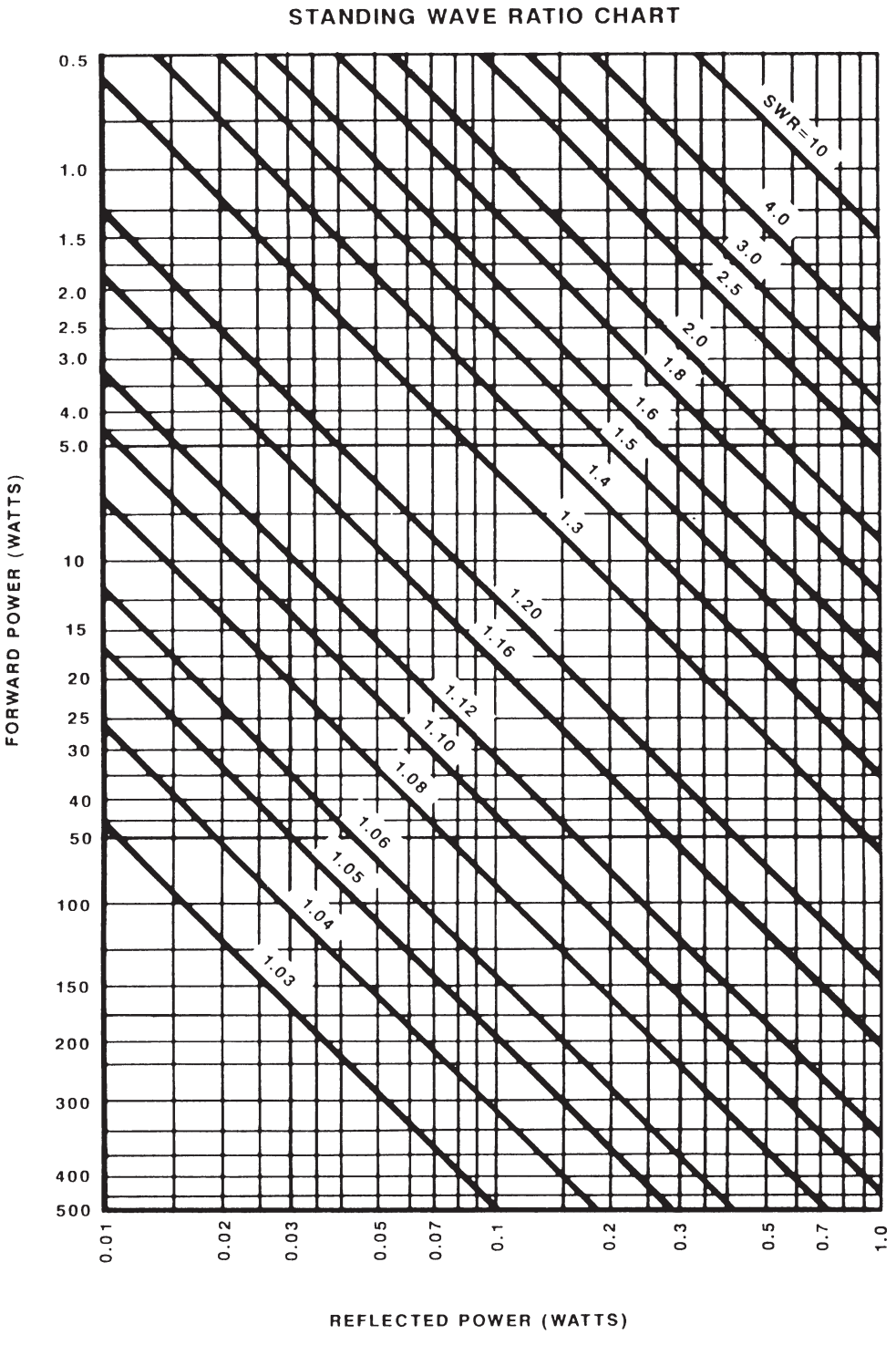


Figure 19-17
Standing wave ratio chart.

SUMMARY

This chapter introduced you to transmission lines, which provide the means of getting radio frequency signals from a source to a load. They are the "pipes" through which radio energy travels.

You studied several types of transmission lines, including balanced lines like twin-lead, and unbalanced lines like coaxial cable. You also learned about waveguides, which are used at very high frequencies and above.

In the study of standing waves on transmission lines, you learned about forward and reflected power, and also about resonant and non-resonant lines. As a part of this, you found out how to use a section of transmission line as a transformer for impedance matching.

SELF-TEST REVIEW

1. What is the primary purpose of a transmission line? _____

   _____

2. What is the difference between a balanced and an unbalanced transmission line? _____

   _____

   _____

   _____

3. The characteristic impedance of a transmission line is determined by its _____ .

4. What are the equations for determining the standing-wave ratio of a transmission line? _____

   _____

5. When a transmission line is terminated in its characteristic impedance, the SWR is _____ .

6. What is the input impedance of a 1/4λ short-circuited transmission line at its resonant frequency? _____

7. 1/4λ and 1/2λ transmission lines can be used as _____ circuits.

8. A reflectometer is used to measure _____ .

9. List the advantages and disadvantages of balanced and unbalanced, or coaxial, transmission lines. _____

_____

_____

_____

_____

CHAPTER 20 — Antennas

INTRODUCTION

After a transmitter generates an RF signal, there must be some method of radiating this signal into space. There must also be some method for a receiver to intercept, or "pick up," the signal. The antenna fulfills both requirements.

An antenna converts high frequency current into electromagnetic waves for transmission, and does just the opposite for reception. Transmitting and receiving antennas have different functions, but they behave identically. That is, their behavior is "reciprocal."

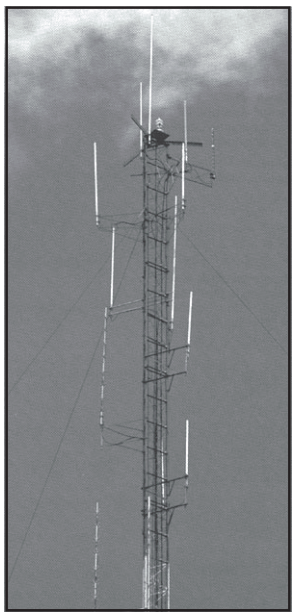


Photo courtesy Andrew Corp.

OBJECTIVES

When you complete this chapter, you will be able to:

- Define antenna, field strength, end effects, and the mirror image principle.
- Identify a half-wave dipole from its voltage and current distribution.
- Identify the radiation pattern for a half-wave dipole.
- List the radiation fields present in an electromagnetic wave and state which field determines the polarization.
- State the input impedance of a dipole.
- Find the length of a half-wave dipole for a given frequency.
- Identify the radiation pattern for a quarter-wave vertical antenna.
- Find effective power when given antenna gain, transmission line loss, and transmitter power output.
- Name the two types of antenna arrays.
- List the elements used in a Yagi antenna.
- Explain how parabolic antennas work, and where they are used.
- List two types of driven arrays.
- Explain the difference between driven and parasitic arrays.

ANTENNA PRINCIPLES

Figure 20-1 shows a 1/4λ open-circuited transmission line. You can see that the forward and reflected waves combine to form a standing-wave pattern on the line. This was described previously, but it wasn't mentioned that not all of the forward energy is reflected by the open circuit. Actually, a small portion of the RF energy escapes from the system and is radiated. This happens because the RF energy traveling toward the open circuit is required to undergo a violent change, a phase reversal, when it reaches the open. A small portion of the forward wave cannot do this, and it actually "shoots" past the open circuit into free space. The amount of waves escaping the transmission line is very small, for two reasons. First, if the surrounding space is considered as a load for the transmission line, a mismatch exists, and very little power is dissipated in this "load." Second, since the two wires are close together and 180° out of phase, the radiation from one will cancel that from the other.

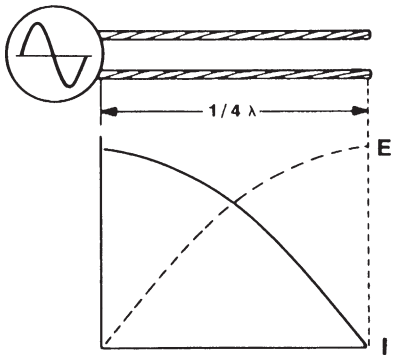


Figure 20-1
Open circuited quarter wave transmission line.

But when the wires are spread apart, the radiation is less likely to cancel. Also, the radiating transmission line is better coupled to the surrounding space because more power is "dissipated," or radiated. Moreover, because the wires are spread out, waves traveling along the line find it increasingly difficult to reverse phase. Thus, everything points to an increase in radiation.

Radiation can be increased even further by spreading the wires until they are in line, as shown in Figure 20-2A on the next page. The electrostatic, or electric field and the electromagnetic field of the line are now fully coupled to the surrounding space. Therefore, the maximum amount of radiation results. This type of antenna is called a

"dipole." The original transmission line was 1/4λ long, but when it is spread out in a dipole, the overall length is 1/2λ. Therefore, the antenna shown in Figure 20-2A is a half-wave dipole.

The current and voltage standing-wave patterns on the half-wave dipole are shown in Figure 20-2. Notice that both ends appear as opens and, therefore, have a voltage maximum and current minimum. The center of the antenna, which is the feed point, has a current maximum and voltage minimum. As a result, the input impedance of the half-wave dipole is low. While the voltage and current distribution indicates that the input impedance should be zero ohms, it is actually 73 ohms. This is due to the energy lost to radiation. This radiated energy is not reflected back to the input and, therefore, complete cancellation does not occur. For this reason, the input impedance rises to 73 ohms as the 1/4λ open transmission line is spread out into a dipole.

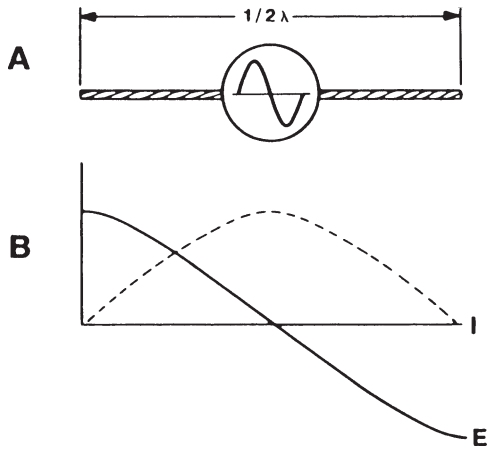


Figure 20-2
The half wave dipole.

ANTENNA RADIATION

Since a dipole has high voltage potentials at both ends, an electrostatic, or electric field exists between these points. This field is just like the one that exists between the plates of a capacitor. However, in the case of the antenna, the electric field is not confined to the area between the plates; instead, it is radiated into space. The electric field around a half-wave dipole is shown in Figure 20-3.

Note also from Figure 20-3 that a magnetic field exists around the antenna. This is due to the current flow in the antenna. Since current is maximum at the center of the antenna, so is the magnetic field. Both the magnetic and electric fields are radiated as an electromagnetic wave. This wave propagates, or travels, outward from the antenna, and will continue traveling even after the current and voltage are removed from the antenna.

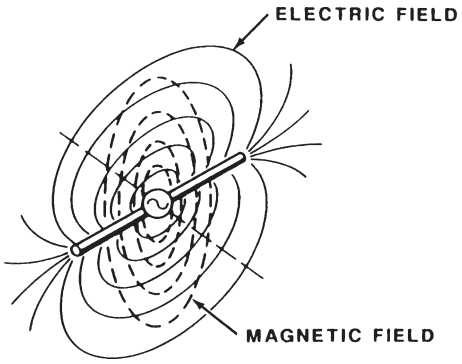


Figure 20-3
Magnetic and electric field around a half wave dipole.

The farther the electromagnetic wave travels from the antenna, the weaker it becomes. The amplitude of this wave, or radiation field, is evaluated in terms of the voltage it will induce across a wire. This is know as its "field strength."

The field strength at any location depends upon the distance from the transmitter and upon the radiated power. The strength varies inversely with distance. For example, if the distance from the antenna is doubled, the field strength is reduced by one-half. Conversely, if the distance is cut in half, the field strength is doubled. Likewise, if the radiated power is increased, the field strength at any location is also increased. However, since field strength is indicated by the voltage induced in a wire, any power increase must be converted to a corresponding voltage increase. Since power is proportional to voltage squared,

$$P = \frac{E^2}{R}$$

It follows that the increase in field strength is proportional to the square root of the power change. Thus, doubling the radiated power will increase the field strength by $\sqrt{2}$ or 1.414. Conversely, to increase the field strength by 2, it is necessary to increase the radiated power by 4.

ANTENNA POLARIZATION

The electromagnetic wave radiated from any antenna has two component fields: the electric field and the magnetic field. These fields are perpendicular to each other and both are perpendicular to the direction of propagation, or travel. This is shown with vectors in Figure 20-4. In this case, the electric field vector is horizontal. Therefore, the wave is said to be "horizontally polarized." If the pattern were rotated 90°, the electric vector would be vertical, and the wave would be "vertically polarized." Thus, the electric field vector determines the polarization of the wave. As you will see later, a horizontal antenna produces horizontal polarization, while a vertical antenna produces vertical polarization. The importance of this is that a horizontally polarized wave induces maximum voltage into a horizontal antenna. Theoretically, a horizontally polarized wave induces zero volts in a vertical antenna. However, in practice, at HF and VHF, this seldom occurs because of wave changes that take place during propagation.

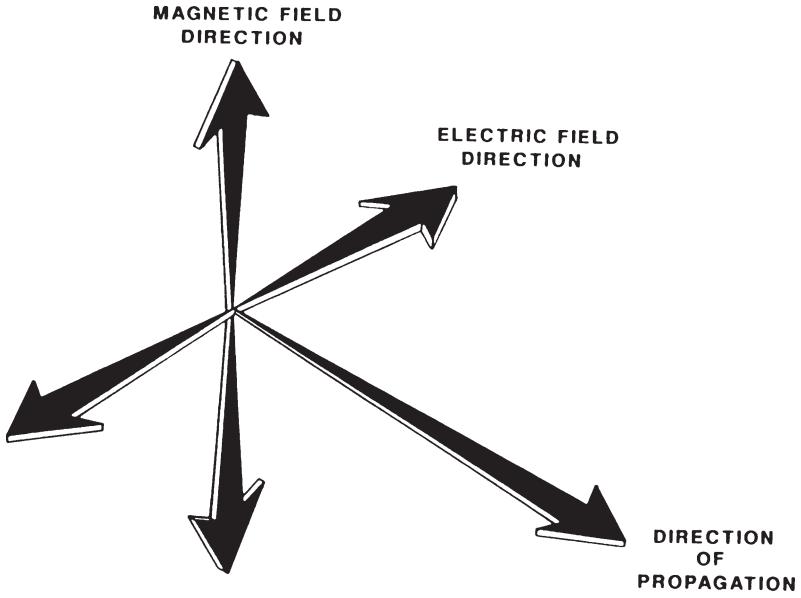


Figure 20-4
Polarization of an electromagnetic wave.

THE DIPOLE ANTENNA

The most commonly used type of antenna is the half-wave dipole. As its name implies, it is 1/2λ long at the operating frequency. The equation for wavelength is:

$$\lambda = \frac{3 \times 10^8}{f}$$

Where λ = wavelength in meters

f = frequency in hertz

This equation gives the wavelength in free space. However, with an antenna, the wave is developed on a wire and end effects must be considered. End effects make an antenna appear electrically as though it were about 5% longer than its actual physical length. This is due to a capacitance that exists between the ends of the antenna. The equation for finding the length of a half-wave dipole that is supported by insulators at each end is:

$$1/2\lambda \text{ in feet} = \frac{468}{f_{MHz}}$$

This equation gives the length in feet and takes end effects into account.

Find the length of a half-wave dipole for 7 MHz.

$$1/2\lambda \text{ in feet} = \frac{468}{f_{MHz}}$$

$$1/2\lambda \text{ in feet} = \frac{468}{7}$$

$$1/2\lambda \text{ in feet} = 66.9 \text{ feet}$$

Therefore, a half-wave dipole for 7 MHz is 66.9 feet long.

A typical dipole installation is shown in Figure 20-5. Notice that the antenna is fed at the center with a 75 ohm balanced transmission line. This gives a good match to the antenna's input impedance.

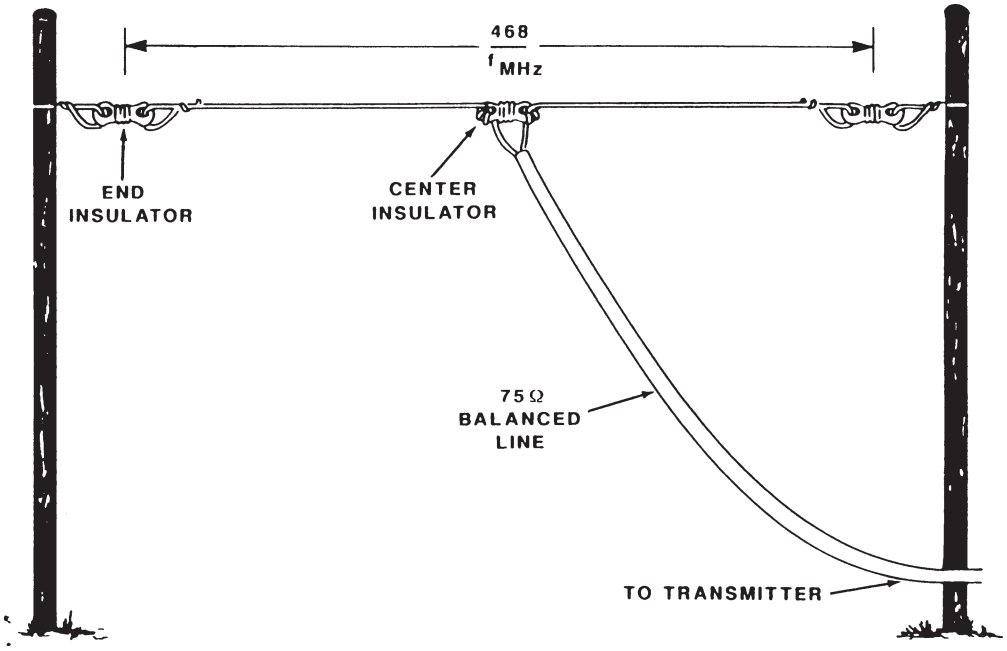


Figure 20-5
Half wave dipole installation.

A dipole can also be "fed" with a coaxial transmission line. The center conductor is connected to one side of the dipole and the shield to the other side. However, since the coaxial line is unbalanced and the dipole is balanced, having been developed from a 1/4λ balanced line, some inefficiency will result. The current and voltage distribution on the antenna will be upset, and RF current will flow on the coaxial shield. This shield current may result in radiation from the cable, which is undesirable.

To correctly couple a coaxial line to a dipole antenna, a balun is used. A balun is a balanced to unbalanced RF transformer. A typical balun is shown in Figure 20-6. This particular balun also acts as a center insulator for the dipole. As shown, the wires connect to the dipole and the connector is for the coaxial cable. Baluns are constructed for specific impedance ratios, just like conventional transformers. When 50 ohm or 75 ohm coaxial cable is used to feed a dipole, an impedance ratio of 1:1 is used.

Figure 20-6
A 1:1 balun.
Photo courtesy Palomar Engineers.

By far, the most important property of an antenna is its pattern of radiation. In the case of a transmitting antenna, this is a graphical plot of the field strength radiated by the antenna in different angular directions. This same radiation pattern also indicates the receiving properties of the antenna. This is because the receiving and transmitting properties of an antenna are reciprocal.

Figure 20-7A shows a 3-dimensional view of the radiation pattern for a horizontal half-wave dipole. Notice that maximum radiation occurs broadside to the antenna and minimum radiation occurs off the ends of the antenna. A polar plot of the radiation in the horizontal plane is shown in Figure 20-7B. The radiation pattern for a vertical half-wave dipole is shown in Figure 20-8A. Note that, in this case, maximum radiation occurs in all directions in the horizontal plane. Minimum radiation occurs directly above and below the antenna. The polar plot for the horizontal plane is shown in Figure 20-8B.

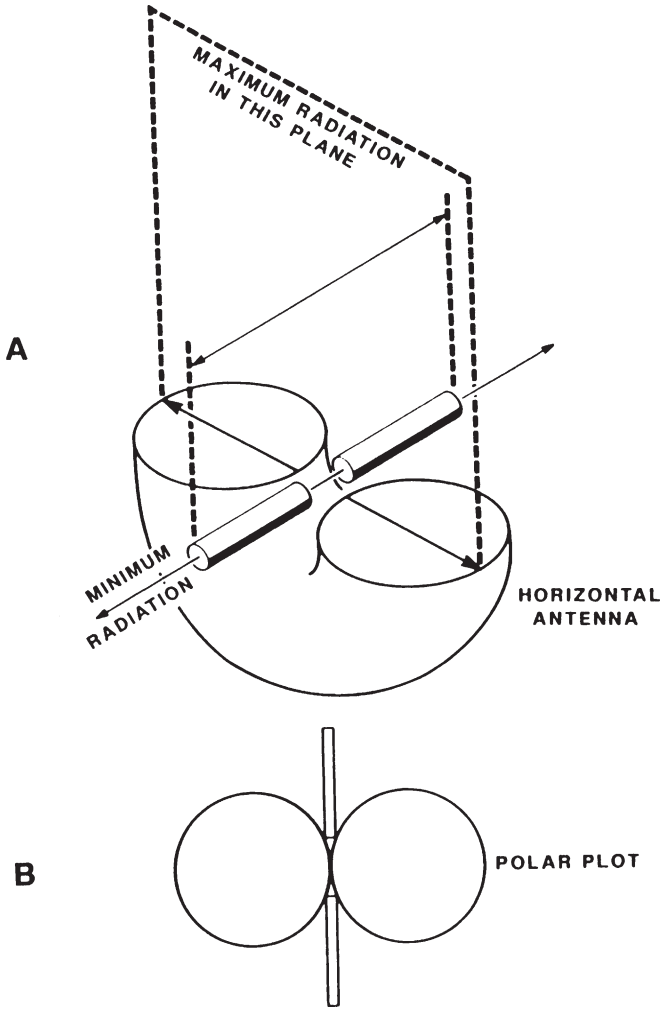


Figure 20-7
Horizontal half-wave dipole radiation pattern.

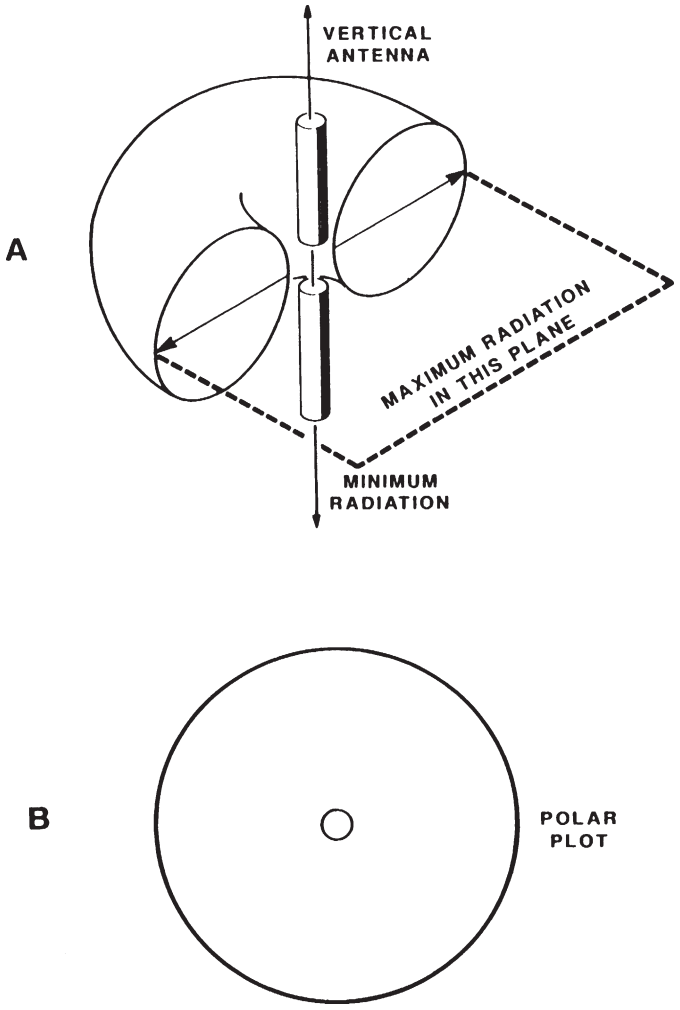


Figure 20-8
Vertical half-wave dipole radiation pattern.

A dipole antenna can also operate on harmonics of its fundamental frequency. That is a half-wave dipole can operate on its second harmonic, where it functions as a full-wave dipole. The current distribution for a full-wave dipole is shown in Figure 20-9A. Note that minimum current occurs at the center of the antenna. This indicates that the center of the antenna is a high impedance point. If a low input impedance is desired, the antenna feedpoint must be moved to either of the current maximums which are located 1/4λ from either end. The radiation pattern for a full-wave dipole is shown in Figure 20-9B. Note that the pattern has split up into four major lobes. This is due to the current distribution on the dipole.

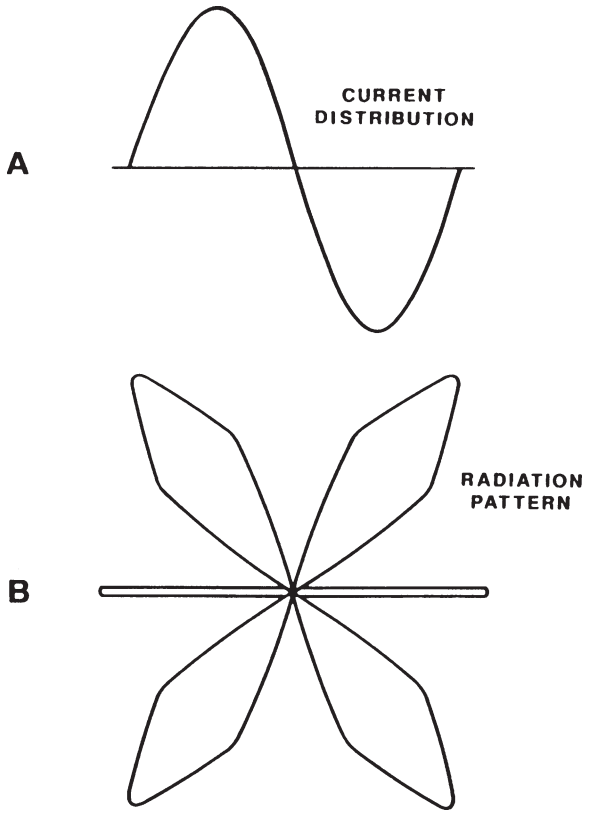


Figure 20-9
The full-wave dipole.

A half-wave dipole can also operate on its third harmonic as shown in Figure 20-10. The current distribution for the resulting 1-1/2λ antenna is shown in Figure 20-10A. Note that, here a current maximum occurs at the center of the antenna. Therefore, the low impedance feedpoint remains at the center of the antenna. The radiation pattern for the 1-1/2λ antenna is shown in Figure 20-10B. Note that there are still four major lobes, but now, there is an additional minor lobe perpendicular, broadside to the antenna.

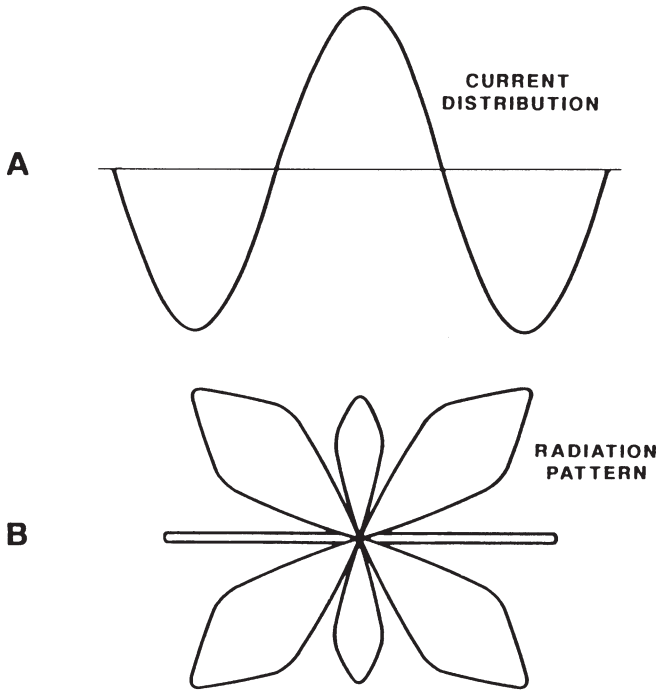


Figure 20-10
1-1/2 wave dipole.

The harmonic operation of dipoles can be put to good use by designing an antenna for use on two harmonically related frequencies.

THE VERTICAL ANTENNA

In cases where vertical polarization is required, the antenna must be vertical. However, at low frequencies, the height of a vertical half-wave dipole can become prohibitive. For example, at 3.5 MHz, a vertical half-wave dipole would be 134 feet high.

This seems to indicate that a low frequency vertical antenna would be impractical. However, if a quarter-wavelength vertical antenna is constructed above a perfect ground, it will have the same characteristics as a half-wave vertical dipole. That's because a perfect ground produces a "mirror image" of the quarter wave, thus, producing the effect of a half-wave antenna. The image results from the reflected radio waves as shown in Figure 20-11A. The voltage and current distribution for a quarter-wave vertical is shown in Figure 20-11B. In order for this antenna to operate correctly, the ground must be a perfect conductor. If it is not, considerable power will be lost in the resistance of the grounding system. When an antenna is constructed over soil of poor conductivity, such as rocky or sandy soil, an artificial ground system must be used. This consists of several 1/4λ copper wires extending outward from the base of the antenna in as many directions as possible. These wires are referred to as "radials" and, usually, four radials are the bare minimum. When only four radials are used, they should be perpendicular to each other.

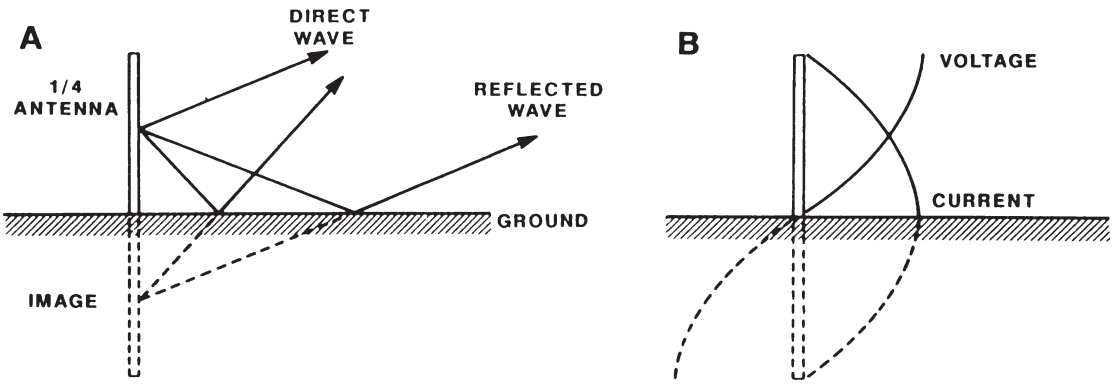


Figure 20-11
The quarter wave vertical.

The radiation patterns for a quarter-wave vertical antenna are shown in Figure 20-12. In the horizontal plane, the antenna radiates equally well in all directions. It is omnidirectional and gives good "coverage" in all directions. In the vertical plane, the radiation is directed low toward the horizon. This gives it excellent long distance propagation characteristics. Wave propagation is described in detail in the next section.

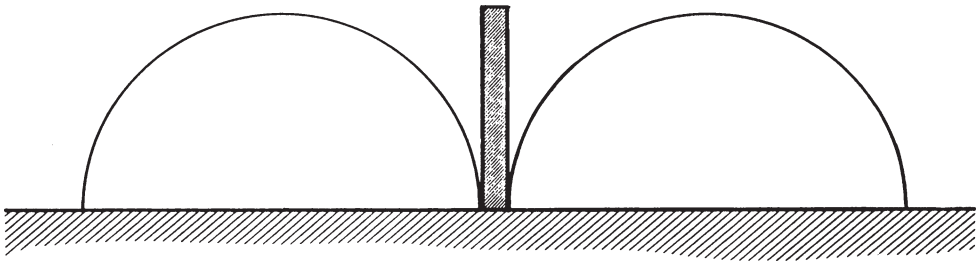


Figure 20-12
Radiation patterns for a quarter-wave vertical antenna.

The input impedance of the quarter-wave antenna is around 36 ohms when it is used with a perfect ground. This is a fair match for 50 ohm coaxial transmission line. Also, since the vertical antenna is unbalanced, because one side is ground, unbalanced coaxial cable is a perfect feeder. The center conductor is connected to the quarter-wave antenna, which is insulated from ground, and the shield is connected to ground, or the radial system.

HORN ANTENNA

When using waveguide, it isn't always practical to build a dipole antenna. Instead, the horn antenna is often used at high frequencies. A horn antenna, shown in Figure 20-13, can be thought of as a transitional piece of waveguide that matches the characteristic impedance of the waveguide to the characteristic impedance of free space which is approximately 377 ohms. The underlying mathematics are complex, but you can intuitively see how a horn works; it provides a gradual transition for the radio energy from the confines of the waveguide to the freedom of the atmosphere.

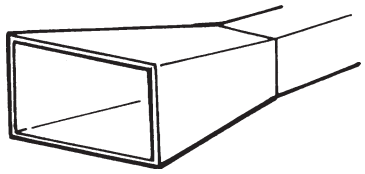


Figure 20-13
A horn antenna.

Horn antennas are widely used at UHF and microwave frequencies, and form the basis of more complex antennas used to increase the directivity of the radiated signal.

ANTENNA ARRAYS

In many situations, the direction of the radiated wave must be restricted within some specific limits. This is done to eliminate or reduce interference to, or from, other stations. It also results in more efficient operation, since any power radiated is directed toward the specific receiving station. Therefore, less power is wasted.

When the radiation pattern of an antenna system is essentially in one direction, it is unidirectional. Contrast this with the half-wave horizontal dipole, which is bi-directional, and the quarter-wave vertical, which is omnidirectional. The measure of a unidirectional antenna's directional properties, or "directivity," is its beam width. This is measured from the radiation pattern as the angle between the two points on either side of maximum radiation where the field strength drops 3 dB. A typical unidirectional radiation pattern is shown in Figure 20-14. In this case, the beam width is 50°.

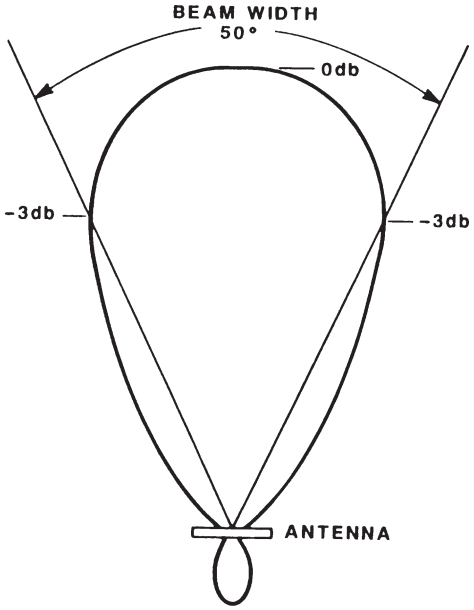


Figure 20-14
Measuring beam width of a unidirectional antenna.

An antenna system with good directivity, or narrow beam width, has the advantage of providing directional gain. Since the radiated power is concentrated within a relatively small beam, the field strength within this beam is much greater than would be obtained from an omnidirectional antenna. Therefore, the effective radiated power (ERP) of the transmitter is increased by the directional gain of the antenna. Directional gain is the ratio of the power required to produce a given field strength at a given location using a reference antenna, compared to the power required to produce the same field strength with a directional antenna. Reference antennas are usually half-wave dipoles or quarter-wave vertical antennas. Very high gains are possible, particularly at VHF and UHF, where electrically large antenna systems can be built in a relatively small space.

Effective radiated power, or ERP, can be calculated if transmitter output power, transmission line loss, and antenna gain are known. For example, if the transmitter output power is 100 watts, transmission line loss is 10 watts, and antenna power gain is 10; ERP is:

100 W	Transmitter output
–10 W	Line loss
90 W	Input to antenna
× 10	Antenna power gain
900 W	Effective radiated power

This shows how a directional antenna greatly increases the effective radiated power.

Directivity, or directional gain, is obtained by using combinations of two or more antenna elements to form an antenna array. Depending upon the method used to excite the additional elements, antenna systems are classified as parasitic or driven arrays.

Parasitic Arrays

An antenna element that is not connected to the transmission line develops a voltage through induction. This element is called a "parasitic element." Figure 20-15 shows a parasitic element located 1/4λ from a dipole, which is called the "driven element." Both elements are 1/2λ long and, therefore, are resonant at the operating frequency. The radiation pattern of the dipole, by itself, is bi-directional and is maximum broadside to the antenna. If you consider the energy traveling toward the parasitic element, it travels 1/4λ. Therefore, before it reaches the parasitic element, it has gone through a 90° phase change. As the wave cuts the parasitic element, a voltage is induced that is 180° reversed with respect to the wave that induced it. As a result, current flows through the element and the element radiates. In the direction beyond the parasitic element, the field radiated by it is opposite to the field from the driven element. Therefore, the two fields cancel, and radiation in this direction is negligible. However, the parasitic element also radiates toward the driven element. By the time this field reaches the dipole, it has gone through a further 90° phase change. Therefore, the total phase change is 360°, and the wave arrives in phase with the energy from the driven element. Thus, radiation toward the driven element is reinforced and maximum radiation results. The direction of maximum radiation is called the forward direction, and the parasitic element is called the "reflector."

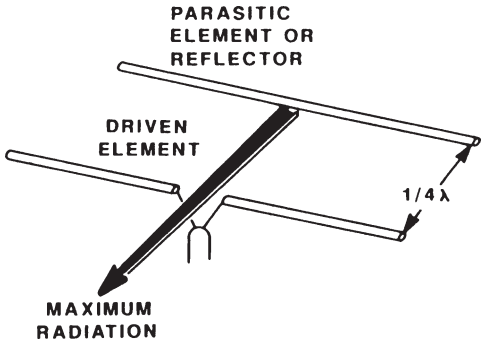


Figure 20-1 5

A half-wave dipole driven element with a parasitic reflector.

The radiation pattern obtained with parasitic arrays depends upon the magnitude and phase of the current in the parasitic elements. These factors depend on the length of the parasitic elements and on the spacing between the driven and parasitic elements. Optimum reflector operation is obtained with a spacing of 0.18λ to 0.2λ and with

an element length approximately 5% longer than 1/2λ. Also, if the parasitic element is approximately 5% shorter than 1/2λ and the spacing is reduced to about 0.1λ, radiation in the direction of the parasitic element is reinforced. In this case, the element is called a "director. "

A 3-element parasitic array is shown in Figure 20-16A. This is known as a "Yagi" antenna, and it is named after one of its inventors. Figure 20-16B shows the radiation patterns of the driven element alone, with the reflector added, and with the director added. You can see that the directivity improves with each additional element. Still sharper directivity and therefore, increased gain can be achieved by adding more directors. An example is shown in Figure 20-17. This is a 6-element Yagi designed for 440 MHz. Notice that it is oriented for vertical polarization.

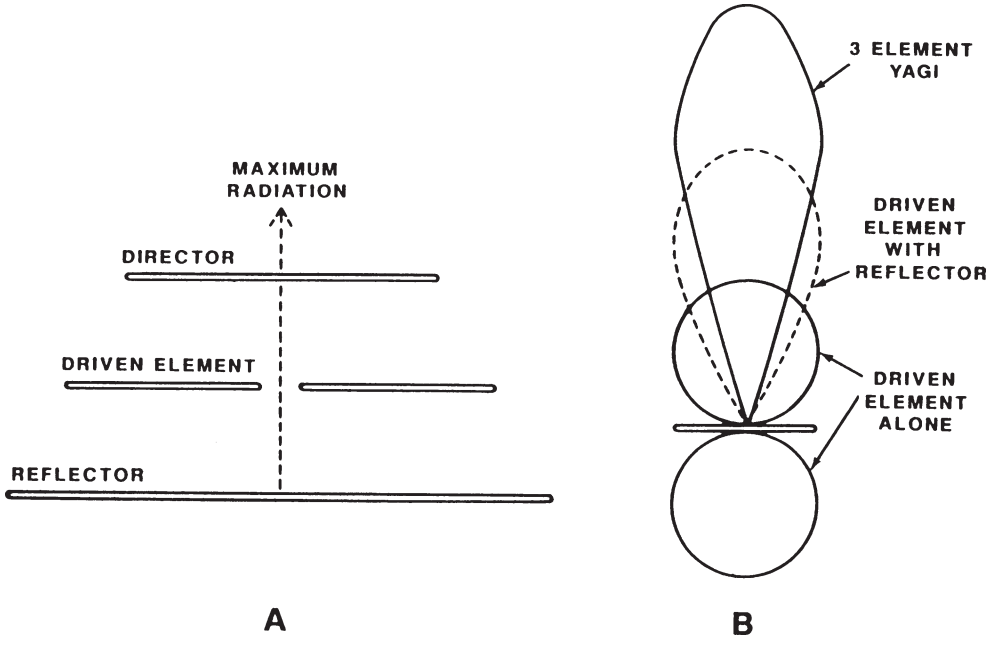


Figure 20-16
A three-element Yagi and its radiation pattern.

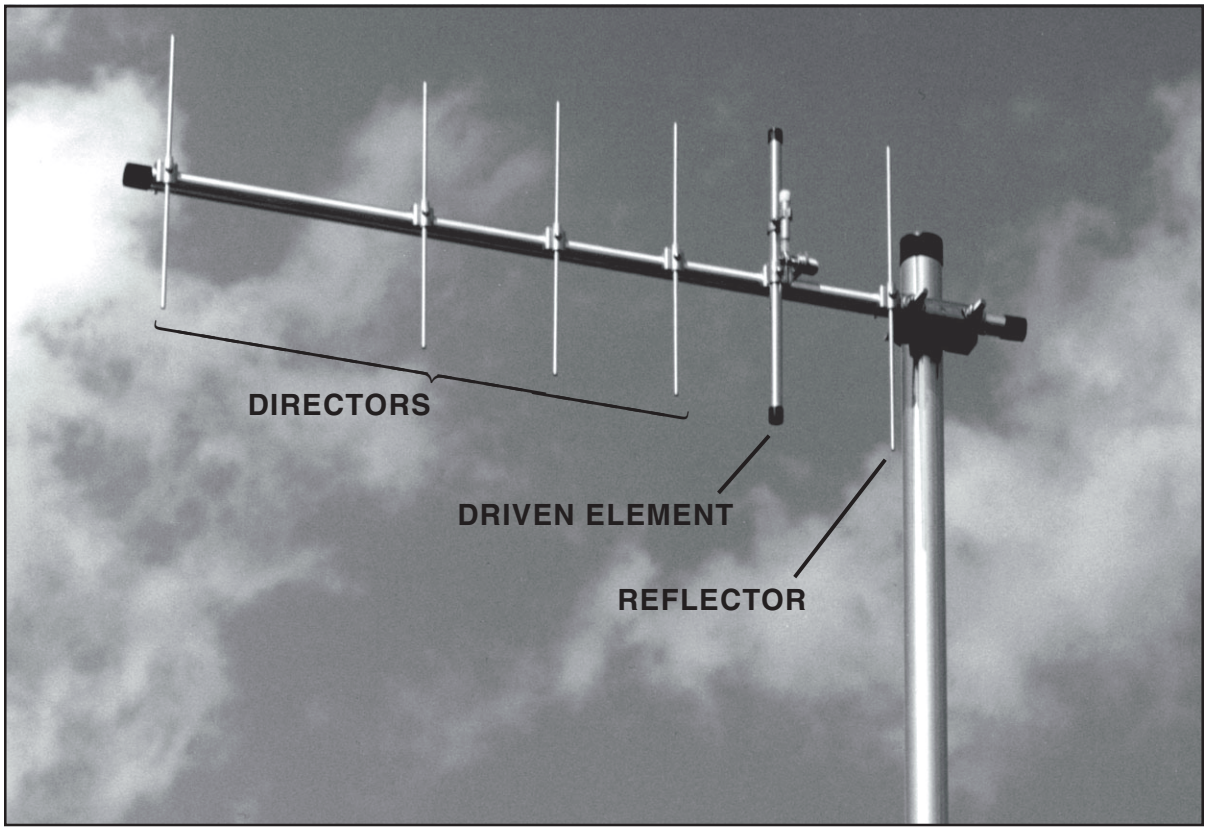


Figure 20-17
A six element Yagi for 440 MHz.
Photo courtesy Cushcraft Corp.

Yagi antennas can provide gains on the order of 15 dB, if they are carefully designed. However, they are narrow-band antennas, and thus, somewhat limited in their application.

The Yagi is not the only type of parasitic array antenna. There are other types of reflectors used to create parasitic antennas, many of them in more frequent use than the Yagi.

One type of antenna that is often used to receive UHF television signals is the corner reflector shown in Figure 20-18. The driven element is, once again, a half-wave dipole. However, the reflector is a corner shaped piece of metal or mesh. There are many configurations of this antenna, but typically the sides of the reflector are twice the spacing between the reflector corner and the driven element. How ever, the dimensions are not critical, and the antenna is considerably smaller than a Yagi array providing the same gain at the same fre quency. In addition, corner reflector antennas are relatively wideband. Depending upon the spacing between the driven element and the reflector, corner reflector antennas are capable of providing 10 to 13 dB of gain using a simple physical structure.

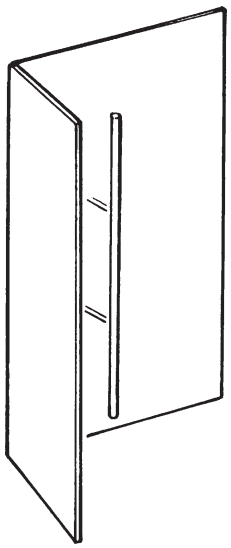


Figure 20-18
Corner reflector antenna.

Another widely used antenna is the parabolic reflector antenna that is often used at microwave frequencies. This is shown in Figure 20-19.

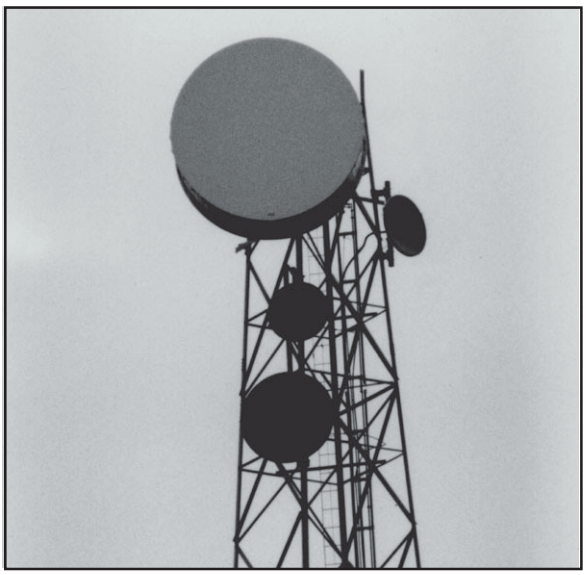


Figure 20-19
Parabolic antennas.
Wheeler Associates Limited photo.

From analytical geometry, you learned that a parabola has a focusing property. Any ray that begins at the focus of the parabola is reflected parallel to the axis of the parabola, as shown in Figure 20-20. This means, that if a driven antenna is placed at the focus of a parabola, then, all of the energy radiated towards the parabola is reflected outward, along the axis of the parabola. Because this property is geometrical, and not electrical, a parabolic reflector should also exhibit relatively wide bandwidth.

Without belaboring the mathematics, it is important to note that parabolic reflectors work just as Figure 20-20 suggests they should. Because their physical sizes are practical at high frequencies, they are widely used in microwave communications. Gains of 35 dB are not uncommon.

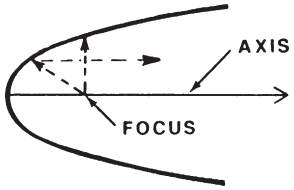


Figure 20-20
Properties of a parabola.

It is possible to obtain satisfactory results using only a portion of the paraboloid surface as the reflector. The less material that is used, the lower the cost of the antenna and its mountings, so it is common to see antennas with only a segment of the parabolic reflector physically in use. Particularly common is the horn paraboloid antenna shown in Figure 20-21. In this antenna, a horn antenna is used to guide waves right up to a section of a parabola, the top of the antenna. A cutout is made in one side of this structure, and the output of the antenna is taken from there.

Figure 20-21
Horn paraboloid antenna.
Wheeler Associates Limited photo.

Parabolic antennas are widely used in microwave and satellite communications. They are physically rugged, relatively simple to construct, and offer high gain at reasonable cost. They are often fed with horn antennas.

There are many types of reflector-based antennas. The ones described are the more common ones. Another means of improving the effective radiated power of a system is to use driven arrays.

Driven Arrays

When all of the elements of an antenna system are fed by the transmission line, it is called a "driven array." An example of a driven array is shown in Figure 20-22A. This is called a "collinear array," since all of the elements are placed end-to-end or in line. Note that the transmission line is attached to each element. Also, the current in each element is in phase, since the transmission line length to each element is equal. The result is a radiation pattern like that shown in Figure 20-22B. The dotted line shows the pattern obtained from a single dipole.The dashed line shows the increased directivity obtained with two elements. Finally, the solid line shows the pattern with four elements. Normally, these collinear arrays are oriented vertically as pictured in Figure 20-23. This gives omnidirectional coverage, but it also directs the energy down towards the horizon and, thus, gives extended coverage at VHF and UHF.

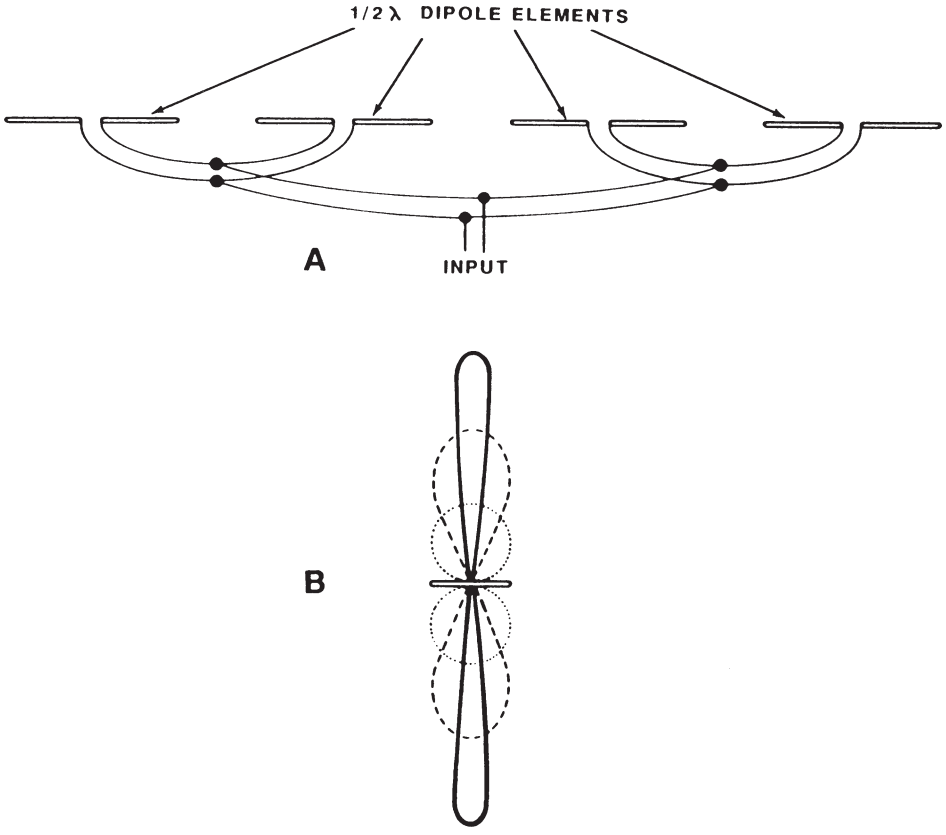


Figure 20-22
A collinear driven array.

Figure 20-23
A four element collinear array.
Photo courtesy Cushcraft Corp.

Another form of driven array is shown in Figure 20-24. This is called a log periodic array, since the element lengths are related logarithmically. Also, each element is fed through a special phasing network. This antenna offers good directivity and gain. However, its greatest advantage is its very wide bandwidth. While most Yagi arrays operate over a very narrow band of frequencies, the log-periodic array can operate over a much wider range of frequencies. For example, the antenna shown in Figure 20-24 gives high gain and an SWR of less than 2:1 from 7.5 MHz to 30 MHz. The best that most Yagi arrays can do is to cover a 1 or 2 MHz band in this same frequency range.

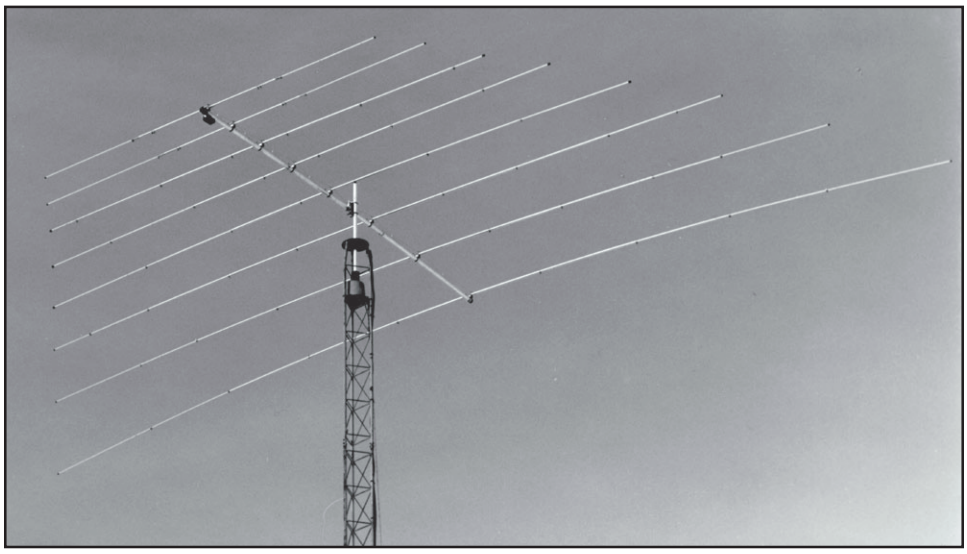


Figure 20-24
A log-periodic array.
Photo courtesy Cushcraft Corp.

SUMMARY

In this chapter you learned about several types of antennas. You found that an antenna is basically a means of matching the transmission line impedance to space, so that the maximum amount of power can be radiated or collected. Some types of antennas have directional characteristics, which are helpful in certain applications.

SELF-TEST REVIEW QUESTIONS

1. A device that converts high frequency current into electromagnetic waves is a/an _____ .

2. The antenna shown in Figure 20-25 is a _____ _____ . The solid line represents the _____ distribution and the dashed line represents the_____ distribution.

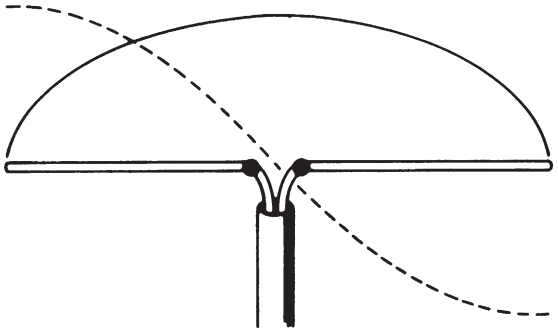


Figure 20-25
Figure for Question 2.

3. What is the input impedance of a half-wave dipole? _____

4. The radiation from an antenna consists of a/an _____ field and a/an _____ field. Both of these fields are radiated as an _____ wave.

5. The amount of voltage induced in a wire by an electromagnetic wave is determined by the wave's _____ .

6. Antenna polarization is determined by the direction of the _____ field vector.

7. An antenna supported by insulators appears approximately 5% longer, electronically, than its actual physical length. This is due to _____ _____ .

8. Find the length of a half-wave dipole for 4 MHz. _____

9. With a half-wave dipole, maximum radiation occurs _____ to the antenna and minimum radiation occurs _____ .

10. The operation of a vertical quarter-wave antenna is based upon the _____ principle and requires a good ground system.

11. A vertical quarter-wave antenna has an _____ radiation pattern in the horizontal plane.

12. If antenna power gain is 5, transmission line loss is 3 watts, and transmitter output is 29 watts, what is the effective radiated power (ERP)? _____

13. The two types of antenna arrays are the _____ array and the _____ array.

14. A 3-element Yagi is an example of a _____ array. Its three elements are the_____, the _____ and the _____ . To increase directivity and forward gain, several _____ can be added to the array.

15. The collinear and log-periodic antennas are examples of _____ arrays.

16. The primary advantage of the log-periodic over a Yagi antenna is its _____ _____ .

17. A horn antenna is used _____ .

18. What are two characteristics of parabolic antennas? _____

_____

_____

GLOSSARY

AGC Circuit — Automatic gain control circuit.

AM/FM Receiver — A radio receiver that receives both amplitude and frequency modulated signals.

Amplitude Modulation — The process of varying the amplitude of a signal with an intelligence signal.

Analog Pulse Modulation — The use of a series of pulses modulated to convey information.

Angle Modulation — Modulation in which the angle of a sine is varied.

Antenna — A device used to radiate radio waves into space or to receive them from space.

Antenna Arrays — A combination of antennas arranged to obtain a desired rejection pattern.

Automatic Gain Control — AGC

Balanced Modulators — A modulator that has an output of the two sidebands without the carrier.

Balanced Transmission Lines — A transmission line having equal impedance from its conductors and earth ground.

Bandwidth — The width of the band of frequencies a circuit will pass.

Channel — 0 to 4 kHz. (1 voice channel).

Noise — Any unwanted disturbance, usually electro-mechanical, within an electronic environment.

Coding — The electro-mathematical representation of information using code.

Complex waveform — A waveform made up of several frequencies or sine waves.

Data Communications — The transferring of information over great distances, usually digitally.

Demodulators — Perform the reverse of modulators by separating the information signal from the carrier.

Digital Pulse Modulation — The digital representation of an analog signal.

Diode Detector — A circuit that uses one or more diodes to provide a rectified output that is proportional to the input.

Diode Modulator — A modulator that produces sum and difference signals from its inputs.

Diode-Ring Balanced Modulator — A balanced modulator that uses diodes as its nonlinear elements.

Dipole Antenna — A straight radiator, usually fed in the center, most commonly a half wavelength.

Distortion — Any type of interference that affects an electronic transmission, caused by many things such as electro-mechanical disturbances or natural occurrences.

Double Conversion Receivers — A receiver that uses two different IF frequencies, a high and a low, taking advantage of the attributes of both.

Double-Tuned Detectors — A discriminator with two secondary transformer windings, one tuned just above the center frequency, and the other tuned just below the center frequency.

DPM Systems — Digital Pulse Modulation.

Driven Arrays — An antenna with all of its elements fed by a transmission line.

Dynamic Range — The ability to receive both very weak and very strong signals without introducing distortion.

Fourier Analysis — The representation of complex waves by a series of sine waves.

Frequency Conversion — The process of converting a signal to another frequency by mixing it with another frequency.

Frequency Modulation — Varying the frequency of a carrier with an intelligence signal.

Frequency-Division Multiplexing — The process of transmitting two or more signals, separated by frequency, over a common path.

FM Transceivers — The combination of an FM transmitter and an FM receiver.

Horn Antenna — Actually a transitional piece of waveguide that matches the impedance of the waveguide to that of free space.

Image Rejection — Measures the receivers ability to reject a strong signal at the image frequency.

Intermodulation Noise — The unwanted mixing of two or more signals within a non-linear device.

IF Amplifiers — A fixed tuned RF amplifier, tuned to the IF frequency.

IF Rejection — Measures a receiver's ability to reject a strong input at its own IF frequency.

Linear Amplifier — An amplifier that increases a signals level without introducing distortion.

Loss — Any unwanted reduction in signal level.

Modems — Modulates and demodulates audio frequencies to transmit and receive data.

Modulation — The process of adding information to a carrier.

Modulation Index — A method of measuring the degree of modulation in an angle modulation system.

Multiplexing — A method by which multiple signals can be transmitted and received over the same channel.

Noise — Any unwanted disturbance, usually electro-mechanical, within an electronic environment.

Noise Figure — A method used to measure noise in equipment or channel elements.

Parasitic Arrays — Antenna elements that are not connected to a transmission line and are dependent upon induction to develop their transmission voltage.

Percent of Modulation — The degree of modulation of a modulated signal.

Phase-Locked Loop Demodulator — A demodulator where the detector is voltage controlled by an internal error voltage.

Phase Modulation — A modulation process where the phase of a signal is varied with an intelligence signal. (A form of angle modulation.)

Phase Relationships — The differences or similarities in phase between the carrier and the sidebands.

Product Detector — A detector whose output is the product of the input signal and the local oscillator signal.

Pulse Amplitude Modulation — Modulation where the amplitude of the represented analog signal varies the amplitude of the pulsed signal.

Pulse Duration Modulation — PTM, where the amplitude of the represented analog signal varies the duration of the pulse.

Pulse Position Modulation — Modulation where the amplitude of the represented analog signal varies the position of the pulse.

Pulse Time Modulation — Modulation where the amplitude of the represented signal actually determines when the pulse will occur.

Ratio Detector — An FM detector whose output is proportional to the ratio of its FM input voltages.

Reactance Modulator — A modulator whose reactance varies with its input, and in turn varies the frequency of the master oscillator.

Resonant Circuits — Circuits designed to operate at a certain frequency.

Sampling — A process by which a modulated waveform is sampled, and only the sample is sent. The signal is then reconstructed from the sample at the receiving end.

Selectivity — Measures a receiver's ability to select a desired signal and reject all others.

Sensitivity — A measure of the amplitude of the weakest signal that can be received.

Series Modulator — A modulator that uses direct coupling to modulate an RF signal, rather than an expensive transformer.

Shannon's Law — A mathematical equation used to determine the capacity of a communications channel.

Sidebands — The frequency bands on both sides of a carrier.

Single Sideband — A form of AM where only one sideband is transmitted.

Sine wave — A periodic waveform, in which the amplitude of displacement at each point is proportional to the sine of the phase angle of the displacement.

Slope Detectors — A discriminator that operates on one slope of a response curve for a tuned circuit.

Space-Division Multiplexing — Multiplexing where a separate path is provided for each signal.

Square wave — A periodic waveform that is made up of a fundamental frequency and an infinite number of harmonics which gives it its square shape.

Standing Waves — Waveforms generated when power is reflected back to the source on a transmission line.

Standing Wave Ratio — The ratio of maximum current to minimum current of maximum voltage to minimum voltage along a transmission line.

Superheterodyne Receivers — An RF receiver where the incoming signal is converted to an IF frequency before they are demodulated.

Suppressed Carrier AM — A form of AM where the carrier is suppressed leaving the two sidebands to carry the information.

Synchronous Detector — A demodulator which has a reference signal equal to the signal to be demodulated.

Time-Division Multiplexing — The process by which, two or more signals are sent over the same path, at the same frequency, but at different times.

Transmission Time — The time it takes a transmission to go from earth to a satellite and back.

Transmitter — An electronic device used to radiate an RF carrier into space.

Tuned RF Receiver — A radio receiver consisting of several amplifiers that are tuned to specific carrier frequencies, in order to block unwanted frequencies.

Varactor Modulator — A modulator that is frequency varied in accordance with its input.

Waveguides — A system of material boundaries capable of guiding electromagnetic waves.

INDEX

A

absorption modulator, 6-5
AGC circuit, 11-13
AGC performance, 11-9
AM characteristics, 6-16
AM waveform, 3-3
amplitude modulation, 1-11, 3-1
analog pulse modulation, 16-1
angle modulation, 4-1, 12-1
antenna, 20-1
antenna arrays, 20-17
antenna polarization, 20-6
antenna radiation, 20-4
atmospheric, 5-7
attenuation, 8-6
autodyne, 10-8
automatic frequency control, 14-4
Automatic gain control, 10-12
AM detectors, 9-1

B

balanced modulator, 7-3
balanced transmission line, 19-3
bandwidth, 1-5, 2-9, 3-15, 4-14, 6-17
bias, 6-5

C

channel, 5-4
channel capacity, 1-5
circuit analysis, 11-11
coaxial transmission line, 19-13
coding, 17-10
collector modulator, 6-6
communications, 1-1
complex waves, 2-1
crystal lattice filter, 8-5

D

delta modulation, 17-12
demodulator, 13-1
deviation ratio, 4-10
digital pulse modulation, 17-1
diode detector, 9-3
diode-ring balanced modulator, 7-4
diode modulator, 6-3
dipole antenna, 20-7
directional gain, 20-19
distortion, 5-8, 11-8
double-tuned detector, 13-4
double conversion receiver, 10-15
driven arrays, 20-25
driver amplifier, 6-8
dynamic range, 11-7

E

effective radiated power, 20-18

F

filter method SSB transmitter, 8-3
Foster-Seeley discriminator, 13-5
Fourier analysis, 2-8
frequency conversion, 8-9
frequency conversion circuits, 10-6
Frequency division multiplexing, 18-5
frequency domain, 2-1
frequency modulation, 4-3
FM broadcast receivers, 14-3
FM transceivers, 15-1

H

half-wave dipole, 20-13
harmonic, 2-4
heterodyned, 10-5
horn antenna, 20-16

I

IC balanced modulators, 7-9
image rejection, 11-6
information transfer, 1-9
information, 1-1
input devices, 5-3
internal noise, 5-7
IF Amplifiers, 10-11
IF Rejection, 11-7
IF selection, 10-9

L

linear amplifiers, 8-12
log periodic array, 20-27
loss, 5-8

M

mechanical filter, 8-8
modulation, 1-10
    envelope, 3-4
    index, 4-10
    percent of, 3-5, 6-12
    phase, 1-10, 4-5
    pulse amplitude, 16-6
    pulse time, 16-8
    pulse code, 17-11
    pulse duration, 16-8
    pulse position, 16-12
MOSFET, 9-9

N

noise, 5-7
    figure, 11-8
noise and distortion, 6-17
nonlinear mixing, 6-4
nyquist rate, 16-5

O

order wire, 18-6
output devices, 5-3

P

parabolic antennas, 20-24
parasitic arrays, 20-19
parasitic elements, 20-19
PCM transmitter and receiver, 17-11
peak, 6-10
peak envelope power, 8-13
percent of modulation, 3-5
phase-locked loop, 13-8
    demodulator, 13-9
phase modulation, 4-5, 12-5
phase relationships, 6-17
phasing method of SSB, 8-11
pi-section, 6-5
PIN diode, 6-5
power and efficiency, 6-16
product detector, 9-9
progressive series modulator, 6-15
pulse amplitude modulation, 16-6
pulse duration modulation, 16-8
pulse position modulation, 16-12
pulse time modulation, 16-8
pulses, 4-15
push-pull circuit, 6-9

Q

quantization noise, 17-6
quantization, 17-3

R

radio, 1-10
    frequency spectrum, 5-6
ratio detector, 13-7
reactance modulator, 12-3
receiver, 5-9
    specifications, 11-3
receivers, 14-1
reference, 4-9
RF choke, 6-8

S

sampling, 16-3
sawtooth wave, 2-6
schematic analysis, 15-3
selectivity, 11-4
sensitivity, 11-3
series modulator, 6-13
Shannon's law, 1-7
sidebands, 3-9, 4-11
sideband filter, 8-3
sideband selection, 8-10
single sideband, 8-1
sine wave, 2-3
slope detector, 13-3
space-division multiplexing, 18-4
square wave, 2-4
squelch, 15-4
SSB Power Ratings, 8-12
standing-wave ratio, 19-9
standing waves, 19-6
stationary voltage, 19-7
superheterodyne receivers, 10-1, 10-4
suppressed carrier AM, 7-1
SWR measurement, 19-18
synchronous detector, 9-6

T

telegraph, 1-9
time-division multiplexing, 18-10
time domain, 2-3
transistor detector, 9-5
transmission, 19-1
transmitter, 5-4
tuned RF receiver, 10-1, 10-3
TWIN-FET balanced modulator, 7-3

V

varactor modulator, 12-4
variable-level quantization, 17-7
vertically polarized, 20-6
vertical antenna, 20-14

W

waveguides, 19-7